01-0

SCHEDULE

Invention TITLE:	System Health Monitoring
Sun Microsystems Reference:	P8871
US Serial No:	10/653,034

INVENTORS:

JAMES E KING
75a Waterloo Road
Wokingham
Berkshire RG40 2JE
United Kingdom

MARTIN P MAYHEAD
Tuscaloosa
Tilford Road
Hindhead
Surrey
GU26 6QY
United Kingdom

SUN REF: P8871US

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICATION PAPERS

OF

JAMES E KING

AND

MARTIN P MAYHEAD

FOR

SYSTEM HEALTH MONITORING

SUN REF: P8871US

SYSTEM HEALTH MONITORING

BACKGROUND OF THE INVENTION

The present invention relates to monitoring system health during initialization of a computer system.

It is known to provide a watchdog function for monitoring system health of a running system. A watchdog function typically functions by means of the operating system outputting signals at regular intervals to a hardware timer that counts between determined values (e.g., from a given value down to an underflow value, or from a given value to an overflow value). The signals that are output are used to reset the counter to the appropriate given value so that the overflow or underflow value is not reached. This is sometimes described as "patting a watchdog timer". In the event that the operating system hangs, or there is some other system failure that means that the signals are no longer output, then the timer will reach the overflow or underflow value as appropriate and this is used to indicate that a fault has occurred. In such a normal operating state, the frequency of the output signals and the length of the period represented by the difference between the given and overflow/underflow count values can be chosen in a predictable manner such that the overflow or underflow of the counter does not occur during normal operation.

In a highly available system, which includes a plurality of less highly available sub-systems, it may be necessary to restart the sub-systems from time to time to maintain overall operation in response to transient faults. Although a conventional watchdog approach can readily be used to monitor the health of the sub-systems where the operating system of those sub-systems is operational, such an approach to health monitoring cannot be used during state transitioning situations, for example at initialization or in abnormal states where the operating system is not operational.

The present invention seeks to address the monitoring of the health of such sub-systems in such situations.

SUMMARY OF THE INVENTION

Embodiments of the invention can enable monitoring of the health of a system module during state transitioning situations, whereby an indication can be provided of a failed transitioning of the system module.

One aspect of the invention embodiment can provide a method of monitoring the health of a system module in a system during state transitioning, wherein the system further includes a monitor module operationally connected to the system module. The method can include the system module outputting a status signal at predetermined system status points during state transitioning. The method can further include the monitor module start a timer on detecting a first status signal and resetting the timer on detecting a subsequent status signal. The timer can further be operable to indicate a failed transitioning of the system module in the event that the timer is not reset within a determined period.

In an embodiment of the invention, therefore, a failed transitioning of a system module, for example a failed restart of a system module can be detected, enhancing the reliability of the overall system.

In an embodiment of the invention, the state transitioning can, for example, by in response to one of starting the system module and shutting down the system module. For example, a signal can be output by the system module for at least one of the following system status points, namely: at power on self test start; at power on self test end; at power on or reset; at an end of initial hardware power up, at the start of booting, at completion of booting, on a shutdown or panic power-off and on a system reset.

In one embodiment of the invention the timer can be reset on detecting each of a set of successive status signals, whereby the timer is operable to indicate a failed transitioning of the system module in the event that the timer is not reset within a respective determined period for each of a plurality of pairs of successive status signals. An initial period for the timer can be determined to exceed an expected maximum time to a subsequent status signal assuming a healthy system module. The monitor module can be operable to record a time for a given pair of status signals on a given initiation of the system and to adapt the determined period for a subsequent system initiation.

Another aspect of the invention can provide a computer system configured to receive a system module and comprising a monitor module to be operationally connected to the system module. The monitor module can be operable to start a timer on detecting a first status signal output by a received system module at one of predetermined system status points during state transitioning of the system module. The monitor module can be operable to reset the timer on detecting a subsequent status signal output by a received system module at another predetermined system status point during state transitioning of the system module. The timer can thereby be operable to indicate a failed transitioning of the system module in the event that the timer is not reset within a determined period.

A further aspect of the invention can provide a system module for a computer system configured to receive said system module and comprising a monitor module to be operationally connected to the system module. The system module can be operable to output status signals at predetermined system status points during state transitioning of the system module, whereby the monitor module is operable to set a time on receipt of a first such status signal and to reset the timer on detecting a subsequent status signal, and whereby the timer is operable to indicate a failed transitioning of the system module in the event that the timer is not reset within a determined period.

Yet a further aspect of the invention can provide instructions operable to control a monitor module for monitoring the health of a system module system during state transitioning. The instructions can be operable to cause the monitor module to start a timer on detecting a status signal output by the system module for a predetermined system status point during state transitioning. The instructions can be further operable to cause the monitor module to reset the timer on detecting a subsequent status signal. The timer can thus be operable to indicate a failed transitioning of the system module in the event that the timer is not reset within a determined period.

The instructions can be provided on a carrier medium. The carrier medium could be a storage medium, such as a solid state, magnetic, optical, magneto-optical or other storage medium. The carrier medium could be a transmission medium such as broadcast, telephonic, computer network, wired, wireless, electrical, electromagnetic, optical or indeed any other transmission medium.

Further aspects and advantages of the invention will become apparent from the following description of particular embodiments.

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments of the present invention will be described hereinafter, by way of example only, with reference to the accompanying drawings in which like reference signs relate to like elements and in which:

Figure 1 is a schematic representation of an architecture of a multiprocessor system for supporting a web site;

Figure 2 is a schematic representation of a racking system incorporating an example of a carrier in the form of a rack-mountable shelf according to a first example;

Figure 3 is a front view of an example of a carrier in the form of a shelf of Figure 2;

Figure 4 is a rear view of an example of the shelf of Figure 2;

Figures 5A, 5B and 5C are schematic perspective views and a schematic exploded view respectively of an example of an information processing cartridge for mounting in the shelf of Figure 2;

Figure 5D is a schematic perspective view of an example of an information processing cartridge for mounting in the shelf of Figure 2;

Figure 6 is a schematic perspective view of an example of a combined switch and service processor module for mounting in the shelf of Figure 2;

Figure 7 is a schematic perspective view of an example of a power supply module for mounting in the shelf of Figure 2;

Figures 8A, 8B and 8C are a schematic plan view and schematic perspective views, respectively, of an example of the chassis and midplane of the shelf of Figure 2;

Figure 8D is a schematic perspective view of a part of a base member of the shelf of Figure 2;

Figures 9A, 9B and 9C are schematic front, top and rear views, respectively, of an example of a midplane of the shelf of Figure 2;

Figure 10 is a schematic cross section view through the shelf of Figure 2;

Figure 11 is a functional block diagram of an example of an information processing subsystem for the information processing cartridge of Figure 5;

Figure 12 is a functional block diagram of an example of an information processing subsystem for the combined switch and service processor module of Figure 6;

Figure 13 is a functional block diagram of an example of a subsystem for the power supply unit of Figure 7;

Figure 14 is a functional block diagram showing the connectivity between the components of the shelf of Figure 2;

Figure 15 is a functional block diagram showing the external connectivity of the shelf of Figure 2;

Figure 16 is a schematic representation of a shelf showing the external connections from the shelf of Figure 2;

Figure 17 is a schematic representation of a rack mounted system comprising a plurality of such shelves;

Figure 18 is a functional block diagram of a further example of an information processing subsystem for the information processing cartridge of Figure 5;

Figure 19 is a schematic representation of aspects of the embodiments of Figures 11 and 12 for use in explaining an embodiment of the present invention;

Figure 20 is a timing diagram showing the timings of various state transitions of a system module;

Figure 21 is a flow diagram illustrating a first example of operation of an embodiment of the invention;

Figure 22 illustrates various phases in operation of an embodiment of the invention; and

Figure 23 is a flow diagram illustrating the operation of a further embodiment of the invention.

While the invention is susceptible to various modifications and alternative forms, specific embodiments are shown by way of example in the drawings and are

herein described in detail. It should be understood, however, that drawings and detailed description thereto are not intended to limit the invention to the particular form disclosed, but on the contrary, the invention is to cover all modifications, equivalents and alternatives falling within the spirit and scope of the present invention as defined by the appended claims.

DESCRIPTION OF PARTICULAR EMBODIMENTS

Embodiments and examples are described hereafter by way of example with reference to the accompanying drawings.

Shown in Figure 1 is an example of an application of a high capacity multiserver system 1 for implementing a network-connected web site such as, for example, an airline reservation system on the World Wide Web.

As shown in Figure 1, an external network 3 (e.g., the Internet) for communicating with a user 5 can be connected to gateways 7 which can be connected to an entry edge server group 9 implemented by a web farm. The entry edge server group 9 forms an interface to the external network 3. The entry edge server group 9 can then be connected by switches 11 and a firewall 13 to a web edge server group 15 that can also be implemented as a web farm as shown in Figure 1. The web edge server group 15 can serve to cache web pages that are readily accessible to users 5 accessing the system 1 from the external network 3, for example for checking flight times, etc. The web edge server group can comprise a number of blade server (BS) shelves and a number of network addressable storage (NAS) shelves for storing critical data. The web edge server group 15 can be further connected by a further firewall 17 to a plurality of application servers 19, which can be responsible for, for example, processing flight reservations. The application servers 19 can then be connected via a further firewall 21 to computer systems 23, 25, for example, e-commerce services including financial services for receiving and processing payment for airline reservations.

As will be appreciated, the server system described above with reference to Figure 1 is only an example of a possible application for a multiprocessor server system. Multiprocessor server systems have many different applications and the

present system is not limited to being applicable for use in only one or a limited number of such applications, rather multiprocessor server systems as described herein are operable for use in many different applications. A non-exhaustive list of such alternative applications includes: e-commerce web server systems; telecommunications network server systems; LAN application and file server systems and remote vehicle control systems.

With reference to Figure 2, there is shown a schematic perspective representation of a rack system 31 as viewed from the front including left and right front uprights 32 and 33 and left and right rear uprights 34 and 35. The uprights can be formed with apertures for receiving shelf fixings (e.g., screws, bolts, clips, etc., for mounting brackets, slides, rails, etc.).

Also shown in Figure 2 is an example of a blade server shelf 41 mounted in the rack system 31. The shelf 41 forms a carrier configured to carry a plurality of information processing cartridges 43 located side by side along the shelf.

The term “shelf” is used herein in a conventional way to describe a structure that is mountable in rack system 31 and is configured to carry one or more components to form at least a part of a rack-mountable system. In the present example, the shelf 41 is three-dimensional, having a height (H), width (W) and depth (D). In the present example, one dimension (hereinafter described as the height, H) is smaller than the other dimensions (hereinafter described as the depth, D, and the width, W) to facilitate mounting of the shelf within the rack system 31. It will be appreciated that although the width and depth are typically constrained by the dimensions of the racking system for which the shelf is designed, there is more freedom as regard the height, subject to taking account of any appropriate standards and packaging considerations.

Each of the information processing cartridges contains at least one processor. Each information processing cartridge in the present example is operable as a server. In the described examples, the information processing cartridges are configured as robust enclosed modules.

In the example to be described in more detail presently, the information processing cartridges, when aligned in the carrier shelf, look like rectangular slabs, or blades. Accordingly, an information processing cartridge can be described as a blade. The information processing cartridges 43 comprise information processing modules enclosed in an enclosure, or housing, so that the information processing modules have the form of cartridges. Also, as the information processing cartridges are to operate as computer servers in the example described in more detail presently, an information processing cartridge 43 can also be described as a server blade. Accordingly, in the context of this example, the terms module, cartridge and blade are used interchangeably.

The illustrated example of a shelf 41 is configured to carry sixteen information processing cartridges 43, each of which is removably mountable in a respective opening 45 in the front of the shelf, whereby the information processing cartridges can be inserted into and removed from the front of the shelf 41 without removing the shelf 41 from the rack system 31.

In the present example, the shelf 41 comprises a three-dimensional, generally rectangular, enclosure, or housing, 47 that is suitable for mounting in generic racking systems including both 4-post and 2-post systems. It can be mounted on fixed rigid rack mounting ears and/or a simple slide/support system. The present example is designed for standard 19"-wide racking (1"=25.4mm) as defined, for example, in the well-known IEC297 and EIA 310 specification standards with height corresponding to the so-called 3U (3 standard unit) height. For mounting such a 3U unit in such a 19"-

wide racking system, with a depth of, say 25" or 30", the enclosure can be arranged with a height of up to about 130.5mm, a width of up to about 445 mm and a depth, including all hardware and fascias, but excluding cable management, of up to about 635mm, with the depth from the front-most point of a fascia to a rear I/O connector panel of a rear mounted Field Replaceable Unit (FRU) of about 610mm. Of course, other examples designed for other racking systems could have different dimensions.

This example of a shelf 41 has a single enclosure, or housing, 47 that houses a number of modular units or subsystems, the majority of which are replaceable in the field and are therefore known as Field Replaceable Units (FRUs). These modular units include the information processing cartridges 43.

The shelf enclosure 47 can be fabricated from sheet material (e.g., from steel sheet) to form a chassis portion 49 that includes a base 51, two sides 53 and 55, a front 57 and a rear 59. The word "front" as used here is merely used as a label herein to refer to the face, or wall 57 of the enclosure that is located at the main access side of the rack system 31 in use when the shelf is mounted therein. Similarly, the words "rear" and "side" are merely used as labels herein to refer to the faces, or walls 59, 53 and 55 that, in use, are located at those respective positions when the shelf is mounted in the rack system 31.

The openings 45 can be formed in the front face 57 for receiving the information processing cartridges 43 and, as will be explained later, apertures can also be formed in the rear face 59 for receiving further FRUs. The enclosure can further include a removable top cover 61 that can be secured to the chassis portion 49 by suitable fastening (e.g., screws). The apertures in the front and rear faces 57 and 59 allow at least some of the FRUs to be inserted into and/or removed from the shelf enclosure 47 via the front or the rear thereof, as appropriate, without removing the shelf from the racking. Access to components mounted in the shelf that are not

accessible via one of the apertures in the front 47 and rear 59 faces can be achieved by removing the shelf enclosure 47 from the racking system 31 and then removing the top cover 61 of the shelf enclosure 47.

Figure 3 is a front view of an example of a shelf 41 for a first example. A plastic front bezel 63 can be provided that fits on the front face 57 (shown in Figure 2) of the chassis 49 of the shelf enclosure 47. The front bezel 63 can be formed as a unitary removable part that spans the whole width and height of the front of the shelf enclosure 47. The front bezel 63 could alternatively include a number of separate components, or moldings. The front bezel can include a peripheral portion 64 that can provide areas for corporate and product branding marks, for identification and numbering for the information processing cartridge and for a bar code label (all not shown). One or more apertures 65 can be formed in the peripheral portion 64 of the bezel 63. The apertures 65 in the bezel can be arranged to align with one or more apertures (e.g. a slot (not shown in Figure 3) in the front face of the chassis. In use, air can pass through the apertures 65 to flow into the shelf enclosure 47 to reach FRUs that are mounted in the shelf enclosure 47 through the rear face 59 thereof. Air flowing through the aperture 65 flows into a plenum chamber 66 (not shown in Figure 3) to flow past the processing cartridges 43 to reach rear mounted FRUs. A central area 67 of the front bezel 63 can be open allowing access to the openings 45 in the front face 57 of the shelf enclosure 47 for insertion and removal of the information processing cartridges 43. Where no active module is mounted in a location for an information processing module, a blanking panel, or filler panel, such as the blanking panel 44, can be located in that location. LED indicators 69 can be mounted on a system indicator printed circuit board (not shown) behind a designated area of the bezel to provide an indication of system status via light guides incorporated into the bezel. A further system indicator board (also not shown) carrying LED indicators can be provided inside the shelf enclosure to be visible from the rear thereof.

As mentioned above, in the present example of a shelf, up to sixteen information processing cartridges 43 can be installed in respective openings 45 in the front face 57 thereof. The number of information processing cartridges 43 actually installed in any installation is dependent upon the system configuration required. Various features relating to the information processing cartridges 43 that are shown in Figure 3 will be described later.

Figure 4 illustrates the rear of the shelf unit of Figures 2 and 3. This shows two different types of FRU 71 and 81 (4 units in total) that have been inserted into respective apertures 72 and 82 in the rear of the shelf enclosure 47. The FRUs shown in Figure 4 include two Combined Switch and Service Processors (CSSPs) 71 and two Power Supply Units (PSUs) 81. Various features shown in Figure 4 will be described later.

Before proceeding with a more detailed description of each of the FRUs 43, 71 and 81 introduced so far and of the construction of the shelf 41, there follows a brief description of an information processing cartridge 43, a CSSP 71 and a PSU 81 with reference to Figures 3, 4, 5, 6 and 7.

Figure 5A provides a perspective view, partly from the rear, of an information processing cartridge 43. Figure 5B provides a perspective view, partly from the front of the same information processing cartridge 43. Figure 5C provides an exploded perspective view of the construction of the information processing cartridge 43. Here is it to be noted that the term "rear" is applied in the context of the position, when installed, of the information processing cartridge, with respect to the shelf 41 (i.e. in this case the "rear" of the information processing cartridge 43 is the innermost part of the information processing cartridge when it is inserted in the shelf 41). Likewise, "front" refers in the present context to the outermost part of the information processing cartridge when it is inserted in the shelf 41.

With reference to Figures 3, 5A and 5B, it will be noted that information processing cartridges are three-dimensional, having a height (h), width (w) and depth (d). If, as in the present example, the information processing cartridges are to be arranged in a one-dimensional array (a row) across the shelf, then efficient packing for the information processing cartridges is achieved where one dimension (here the width, w) is smaller than the other dimensions (here the depth, d, and the height, h). In a particular example, the enclosure of a processing cartridge 43 has a height h, width w and depth d of 115mm, 26mm and 315mm, respectively, although of course, other examples may have different dimensions.

It will be noted that an enclosure 101 of the present example of an information processing cartridge 43 has six, generally rectangular, faces. For convenience only, the face that is visible from the front of the racking when an information processing cartridge 43 is mounted in the shelf 41 is known as the front face 102. The opposite face is known as the rear face 103. In the present example these two faces, as well as top and bottom faces 104 and 105, and the side faces 106 and 107 have the shape of elongate rectangles.

Although in the present example the information processing cartridges have six generally rectangular faces, it will be appreciated that other examples could have other configurations. For example, rather than having a generally rectangular shape, the side faces of another example of an information processing cartridge could have the general shape of a triangle (whereby the information processing cartridge may then only have five faces), a pentagon (whereby the information processing cartridge may then have seven faces), and so on. Indeed, one or more or all of the edges could be curved. However, it will be appreciated that the present configuration provides advantages for example, in terms of manufacturing, engineering and packing density within a shelf 41.

In this example, the information processing cartridge enclosure 101 is fabricated from two housing portions. The first housing portion 109 can be fabricated from, for example, a metal (e.g., pressed steel) and can include one side face 106, the rear face 103 and part of the top face 104. The part of the top face formed by the metal portion 109 is given the reference 1041 in Figure 5A. The second housing portion 110 can be fabricated from, for example, a plastics material and can include the other side face 107 and the front faces 102 and bottom face 105 and the remaining part 1042 of the top face 104. In the present example, a plastics material used is PolyCarbonate Acrylonitrile Butadiene Styrene (PCABS), however many other plastics materials such as other ABSs and nylons may be used. The metal portion 109 also includes a part 1021 corresponding to the front face 102 and a part 1051 corresponding to a part of the bottom face 105. The parts are covered by the plastics portion 110 when enclosure 101 is fully assembled. These parts are visible in the exploded view of Figure 5C. The two housing portions 109 and 110 are secured to one another by fixings such as screws 118.

As shown in Figure 5A, a groove 108 runs along the length of the plastic top face portion 1042. This groove 108 is provided for interfacing with a guide member of the shelf chassis (not shown in Figure 5A, but see Figures 8B and 8C). A similar groove 113 is formed in the bottom face 105 of the plastics portion 110 as shown in Figures 5B and 5C. This groove 113 is provided for interfacing with a guide member of the shelf chassis (not shown in Figures 5B and 5C, but see Figure 8D).

The provision of the enclosure 101 means that the information processing cartridge 43 can safely be handled by an operator who is not a skilled technician. Also, through the use of the enclosure 101, the information processing cartridge is a robust unit that protects its inner workings from the outside environment and vice versa. The use of a conductive enclosure, e.g., a metal enclosure, can result in the

information processing cartridge including its own electromagnetic shielding. To this end, the enclosure 101 can also be provided with an internal electromagnetic interference (EMI) shielding member 116 as shown in Figure 5C. The EMI shielding member can be located between the internal components 112 of the information processing cartridge 43 (not described in detail with reference to Figures 5A to 5C, but see Figures 11 and 18 below) and the plastics portion 110. The EMI shielding member can be secured to the plastics portion 110, for example by heat bonding or an adhesive. In other examples, the plastics portion could have a conductive layer deposited thereon, or conductive plastics material could be used. In this example, the EMI shielding member 116 is provided with electromagnetic interference (EMI) fingers 114 to ensure good contact with the shelf chassis and the adjacent components. These fingers 114 extend through EMI finger apertures 1141 in the plastics portion 110.

As shown in Figures 3, 5B and 5C, the information processing cartridge 43 incorporates an injector/ejector handle 111 on the front face 102 for facilitating insertion and latching of the information processing cartridge 43 within an aperture in the shelf 41.

As shown in those Figures, the handle 111 of the injector/ejector lever extends substantially the whole height of the front face of the information processing cartridge 43, thereby increasing the mechanical advantage and facilitating injection and ejection of the information processing cartridge 43. As is further shown in those Figures, the front face 102 of the information processing cartridge 43 has perforations 115, in the present example slits, to allow for airflow into the information processing cartridge 43. The front face part 1021 of the metal portion 109 has perforations corresponding to those in the front face 102 such that airflow into the information processing cartridge 43 is not impeded. It will be noted in Figures 3, 5B and 5C that the handle 111 is narrower in its middle than at its ends. This reduces any masking effect of the

handle 111 on the airflow to the perforations 115 in the front face of the information processing cartridge 43 and facilitates grasping of handle 111. The handle 111 can be bowed out from the front face of the information processing cartridge to further facilitate grasping thereof and to reduce any masking effect with regard to the perforations. In other examples, the handle 111 could have an open frame-like structure to further facilitate airflow.

As shown in Figure 5A, the rear face 103 of the information processing cartridge 43 also has perforations 117 to allow for air to be exhausted from the rear of the information processing cartridge 43. A fan can be located within the enclosure 101 of an information processing cartridge 43. In the present example the fan is a combined radial fan (blower) and heat sink to direct cooling air onto a processor of the information processing cartridge, which fan also encourages air movement within the enclosure 101. LED indicators 119 (see Figure 3 and 5B) can be provided on the front face 102 of an information processing cartridge to indicate whether power is on, whether service intervention is required and whether the information processing cartridge 43 can be removed. A connector 120, for example a 40 way single connector attachment (SCA-2) connector (a small computer systems interface (SCSI) connector), can be provided at the rear of the information processing cartridge 43 for electrical connection of the information processing cartridge 43 within the shelf 41. The connector 120 is advantageously able to withstand repeated removals and insertions of the enclosure 101 from and into a shelf 41. The connector arrangement can include a guide pin arrangement to prevent module misalignment during insertion of the information processing cartridge into the receiving location.

Thus an example of features and the construction of an information processing cartridge enclosure has been described. Although particular materials and constructions have been described, it will be appreciated that other examples could be employed. Indeed, it will be appreciated that this particular example relates to only a

possible form for the processing cartridge 43. An example of an alternative construction will now be described with reference to Figure 5D.

Figure 5D provides a perspective view, partly from the rear, of an alternative information processing cartridge 43. Here is it to be noted that the term "rear" is applied in the context of the position, when installed, of the information processing cartridge, with respect to the shelf 41 (i.e. in this case the "rear" of the information processing cartridge 43 is the innermost part of the information processing cartridge when it is inserted in the shelf 41).

In this example, the information processing cartridge enclosure 101 is fabricated from pressed steel to form two chassis portions. The first portion 234 includes one side face 107, and part of each of the front and rear faces 102 and 103 and the top and bottom faces 104 and 105. The second portion 235 includes the other side face 108 and the remaining part of each of the front and rear faces 102 and 103 and the top and bottom faces 104 and 105. The two chassis portions 234 and 235 meet at a groove 236 and are secured to one another by fixings (e.g., one or more screws, not shown). Grooves 236 run along the top and bottom faces 104 and 105 of the enclosure 101 and are provided for interfacing with guide rails of the shelf chassis 49 (not shown in Figure 5, but see Figure 8C). A cover portion that is secured to the chassis portion forms the other side face 106. It will be appreciated however, that in another example, the chassis portions could be joined at a position other than the groove 108, with the groove 108 being formed entirely in one of the chassis portions. Alternatively, the enclosure 101 may be constructed from a number of sheets of steel, with each sheet forming one of the faces.

Figure 6 provides a perspective view, partly from the front, of a Combined Switch and Service Processor (CSSP) cartridge (or CSSP) 71. Here is it to be noted that the term "front" is applied in the context of the position, when installed, of the

CSSP cartridge 71, with respect to the shelf 41 (i.e. in this case the "front" of the CSSP cartridge 71 is the innermost part of the CSSP cartridge 71 when it is inserted in the shelf 41).

With reference to Figures 4 and 6, it will be noted that a CSSP cartridge 71 is three-dimensional, having a height (h), width (w) and depth (d). In a particular example, the enclosure of a CSSP 71 has a height h, width w and depth d of 43mm, 202mm and 278mm, respectively, although of course, other examples may have different dimensions.

An enclosure 121 of present example of a CSSP cartridge 71 has six, generally rectangular, faces. For convenience only, the face that is visible from the rear of the racking when a CSSP cartridge 71 is mounted in the shelf 41 is known as the rear face 122. The opposite face is known as the front face 123. In the present example these two faces, as well as side faces 126 and 127 have the shape of elongate rectangles. The top and bottom faces 124 and 125 are also rectangular, but not elongate in the manner of the front, rear, top and bottom faces. Although in the present example the CSSP cartridges have six generally rectangular faces, as for the information processing cartridges 43 it will be appreciated that other examples could have other configurations.

In this example, the CSSP enclosure 121 is fabricated from steel sheet to form a chassis portion that includes the bottom face 125, the front and rear faces 122 and 123 and the side faces 126 and 127. A cover portion that is secured to the chassis portion forms the other top face 124. The cover portion is secured to the chassis portion by suitable fixings, for example one or more screws 128. It will be appreciated however, that in another example, other faces, or portions, of the enclosure could form the chassis and the cover portions. The provision of the enclosure 121 means that the CSSP cartridge 71 can safely be handled by an operator

who is not a skilled technician. Also, through the use of the enclosure 121, the switch cartridge is a robust unit that protects its inner workings from the outside environment and vice versa. The use of a conductive enclosure, e.g., a metal enclosure, means that the CSSP cartridge includes its own electromagnetic shielding. To this end the CSSP enclosure 121 is provided with EMI fingers 129 to ensure good contact with the shelf chassis and the adjacent components.

As shown in Figure 4, the CSSP cartridge 71 incorporates two D-shaped handles 132 to facilitate insertion and removal of the CSSP cartridge 71 with respect to an aperture 72 in the rear face of the shelf enclosure. A latch member 131 can be pivotably mounted on a plate that can be secured (e.g., using screws) to the rear face of the shelf enclosure. The latch member 131 is configured to engage one of the handles 132 and to secure the CSSP cartridge 71 in place. In other examples, the CSSP cartridge 71 could be provided with an injector/ejector handle in a manner similar to the information processing cartridge. As shown in Figure 6, the front face 123 of the CSSP cartridge 71 has perforations 133 to allow for airflow into the CSSP cartridge 71. As shown in Figure 4, the rear face 122 of the CSSP cartridge 71 has perforations 135 to allow for air to be exhausted from the rear of the CSSP cartridge 71.

At least one fan can be located, for example behind the perforated portion 135 of the rear face, in a CSSP cartridge 71 to channel cooling air through the CSSP cartridge 71 from the front to the rear. In this particular example shown, two fans are provided, one behind each set of perforations 135. LED indicators 137, as shown in Figure 4, can be provided on the rear face 122 of the CSSP enclosure 121 to indicate whether power is on, whether service intervention is required and whether the switch can be removed. Additional link status indicators can be provided integral to 2x4 stacked RJ-45 connectors 139, also shown in Figure 4. As shown in Figure 4, electrical connections 141 can be provided at the front face of the CSSP (i.e. on the

face that in use is inside the shelf enclosure 47). Suitable connections for use in the present example include a connector for power connections, a connector for serial management data connections and a connector for information connections. In the present example, information connections are implemented using an Ethernet information communication protocol, e.g. at 1 Gigabit (Gb). However other protocols could equally be used, for example the Infiniband information communication protocol. The connector arrangement can include a guide pin arrangement to prevent module misalignment during insertion of the CSSP module into the receiving location. For this purpose, guide pin holes 142 can be provided on the front face 123 into which guide pins may pass to aid module alignment.

In the present example, up to two CSSPs 71 can be mounted at any one time at the rear of the shelf unit in corresponding apertures 72 in the rear face of the shelf enclosure 47. The number of CSSPs 71 provided in any particular implementation depends upon system configuration and the need, or otherwise, for redundancy.

It will be appreciated that one possible construction of the CSSP cartridge 71 has been described and that as for the information processing cartridge 43, other examples could employ other materials and/or constructions.

Figure 7 provides a perspective view, partly from the front, of a power supply unit (PSU) cartridge 81. Here is it to be noted that the term "front" is applied in the context of the position, when installed, of the PSU cartridge 81, with respect to the shelf 41 (i.e. in this case the "front" of the PSU cartridge 81 is the innermost part of the PSU cartridge 81 when it is inserted in the shelf 41).

With reference to Figures 4 and 7, it will be noted that PSU cartridge 81 is three-dimensional, having a height (h), width (w) and depth (d). In this particular example, the order to provide for a dense packing of the FRUs in the rear of the shelf

41, the PSU cartridge 81 has two dimensions (hereinafter described as the width, w, and the depth, d) that are generally similar. In a particular example, the enclosure of a PSU cartridge 81 has a height h, width w and depth d of 83mm, 202mm and 276mm, respectively, although of course, other examples may have different dimensions.

An enclosure 145 of present example of a PSU cartridge 81 is of generally oblong shape, but has the "top" "front" edge cut away to form an additional "top" "front" sloping face. The enclosure 145 therefore has five, generally rectangular, faces and two faces of generally rectangular shape with one corner cut away. For convenience only, the face that is visible from the rear of the racking when the PSU cartridge 81 is mounted in the shelf 41 is known as the rear face 146. The opposite face is known as the front face 147. In the present example these two faces and the two side faces 150, 151 are of elongate, generally rectangular shape with one corner cut away, given that the width and depth of the PSU cartridge are similar, whereas the top and bottom faces 148, 149, although still rectangular, are not, in this example, notably elongate. A top front face 148a is present at the top front of the enclosure. Thus the front of the enclosure is sloped at the top edge. As for the information processing cartridges 43, however, it will be appreciated that other examples could have other configurations.

In this example, the PSU cartridge enclosure 145 is fabricated from steel sheet to form a housing portion that includes the bottom face 149, the side faces 150 and 151 and the front and rear faces 146 and 147. Cover portions that are secured to the housing portion form the top face 148 and top front face 148a. The cover portions are secured to the chassis portion by suitable fixings, for example one or more screws 152. It will be appreciated however, that in another example, other faces, or portions, of the enclosure could form the chassis and the cover portions. The provision of the enclosure 145 means that the PSU cartridge 81 can safely be handled by an operator who is not a skilled technician. Also, through the use of the enclosure 145, the PSU

cartridge 81 is a robust unit that protects its inner workings from the outside environment and vice versa. The use of a conductive enclosure, e.g., a metal enclosure, means that the PSU cartridge includes its own electromagnetic shielding. To this end the PSU enclosure 145 is provided with EMI fingers 153 to ensure good contact with the shelf chassis and the adjacent components.

As shown in Figure 4, the PSU cartridge 81 incorporates two D-shaped handles 156 to facilitate insertion and removal of the PSU cartridge 81 with respect to an aperture 82 in the rear face of the shelf enclosure. A latch member 155 can be pivotably mounted on a plate that can be secured (e.g., using screws) to the rear face of the shelf enclosure. The latch member 155 is configured to engage one of the handles 156 and to secure the PSU 81 in place. In other examples, the PSU 81 could be provided with an injector/ejector handle in a manner similar to the information processing cartridge. As shown in Figure 7, the front face 147 of the PSU cartridge 81 has perforations 157 to allow for airflow into the PSU cartridge 81. As shown in Figure 4, the rear face 146 of the PSU cartridge 81 also has perforations 159 to allow for air to be exhausted from the rear of the PSU cartridge 81.

A pair of fans can be located behind the perforated portions 159 of the rear face of a PSU cartridge 81 to channel cooling air through the PSU cartridge from the front to the rear. LED indicators 161 can be provided on the rear face 146 of the PSU enclosure 81 to indicate whether input power is good, whether output power is good, whether service intervention is required and whether the PSU can be removed. Electrical connectors 163 can be provided at the front face of the PSU (i.e. on the face that in use is inside the shelf enclosure 47) for connection to the shelf. The PSU 81 of the present example may suitably employ an SSI-MPS (Server Systems Interface - Midrange Power Supply) compliant right angle connector at the front face 147 of the PSU 81 to connect to the shelf 41. The power inlet 83 for each PSU 81 can incorporate a cable/connector retention mechanism (not shown) on the rear face 146

of the PSU to prevent accidental or malicious removal of the power input cord from the PSU 81.

In the present example, the shelf unit enclosure provides slots at the rear of the shelf unit for two hot-swappable, AC input PSUs 81. To provide redundant power supply, both PSU slots are populated. More generally, N+M power supplies can be provided, where N is the minimum number of power supply units needed to support the components mounted in the shelf, and M is a selectable number of 0, 1 or more power supplies to provide for redundancy. In the present example, N = M = 1.

It will be appreciated that one possible construction of the CSSP cartridge 71 has been described and that as for the information processing cartridge 43, other examples could employ other materials and/or constructions.

If the full complement of information processing cartridges or switches are not fitted to the enclosure, then blanking panels/modules (e.g., the blanking panels 44 shown in Figure 3) are fitted to all unpopulated positions to maintain the correct airflow and thermal characteristics of the shelf, a safe internal operating temperature for the shelf life expectancy of all shelf components, electromagnetic compliance (EMC) containment and electrostatic discharge (ESD) containment.

It will be noted that each of the FRUs, such as the information processing cartridges 43, is advantageously contained in its own robust enclosure to facilitate EMC containment, ESD containment, handling, storage and transportation. Each FRU can be configured as a 'sealed' unit in the sense that it can be configured not to have field or customer serviceable parts internally. The FRUs can be configured readily to plug into the shelf enclosure and to be hot swappable. The FRUs can be keyed to prevent incorrect positioning and insertion into the shelf enclosure and are arranged positively to be retained in the shelf by a latching/locking mechanism.

The examples of FRUs described above are not provided with removable media. In the present example, internal data storage is provided by 2.5" IDE 9.5mm or 12.7mm profile hard disk drive (HDD) devices mounted internally in each information processing cartridge 43 and in the CSSP cartridge 71. The drives are not considered as FRUs and are not hot-swappable disk drives in the present example, although they could be in other examples. In other examples, the information processing cartridges can be configured without internal hard disk drives.

The internal configuration of the shelf 41 and a midplane 171 contained therein is described in the following with reference to Figures 8A, 8B 8C and 8D, and Figures 9A, 9B and 9C.

Figure 8A is a schematic plan view showing the internal configuration of an example of a shelf 41 with the cover 61 removed. Figure 8B is a schematic perspective view from above the rear of the chassis portion 47 of the shelf enclosure with the field replaceable units removed. Figure 8C is a schematic perspective view from below the front of the chassis portion 47 of the shelf enclosure with the field replaceable units and the base 51 removed. Figure 8D is a schematic perspective view from the front and above a part of the base 51 of the shelf 41. Figures 9A, 9B and 9C are, respectively, front, top and rear views of the midplane 171. In this example, the midplane is, in use, mounted vertically within the shelf 41 extending across the width W of the shelf 41 at a position approximately half way between the front and the rear of the shelf 41.

The vertically mounted midplane 171 extends, in this example, across the shelf 41 and allows for the electrical interconnection of the FRUs. The various apertures in the front and rear faces 57 and 59 of the shelf 41, in combination with the midplane 171, can be provided with guides (e.g., rails 181) and keying e.g., offset connector

positioning for the insertion of the FRUs into the enclosure and midplane 171. The midplane 171 can be a double-sided, or multi-layer printed circuit board (PCB) assembly that can be mounted vertically in a rigid manner within the enclosure. It can carry connectors 175 on a front surface 172 for making electrical connection with corresponding connectors 120 on the information processing cartridges 43. It can also carry connectors 177 and 179 on rear surface 173 for making electrical connection with corresponding connectors 141 and 163 on the CSSPs 71 and the PSUs 81, respectively. Conductive tracks (not shown) on and through the midplane 171 can be provided to interconnect the various connectors. In addition, the midplane can provide connectors for receiving corresponding connectors connected to first and second indicator boards 183 and 184 that each carry a respective set of LED indicators 69. In the present example, the midplane 171 is not configured as a FRU and is not hot swappable. It is perforated to facilitate airflow through the shelf 41. The midplane 171 can include openings 185, which cooperate with openings in the enclosures of the FRUs 43 and 81, to provide a path for cooling air to pass from the front to the rear of the shelf 41, the cooling air being driven by fans in one or more of the FRUs, for example in the PSUs 81, possibly also in the information processing cartridges 43.

A plenum chamber floor member 94 can extend horizontally from the front of the midplane 171 to the front face 57 of the shelf enclosure, or chassis 47. The member 94 provides a floor for a plenum chamber 66, which is supplied with air via the apertures 65 in the front bezel and, in the illustrated example, the slot shaped aperture 68 in the front face 57 of the shelf enclosure 47. Although, for reasons of ease of illustration a slot shaped aperture 68 is shown, a plurality of apertures 68 aligned with the blade receiving locations may be provided. The aperture or apertures 68 can serve both as air vents for a flow of air to the plenum chamber 66, and also as latching locations for latching portions at the top of the injector/ejector levers 111 for

the blades shown in Figures 5B and 5C. The top and sides of the plenum chamber are provided by the top cover 61 and side faces 53 and 54 of the shelf enclosure 47.

A plurality of cartridge guides 97 can be provided at the underside of the plenum chamber floor member 94. In the present example, these guides comprise sprung wire members, e.g., of a resilient metal such as spring steel, that are attached to the top surface of the plenum chamber floor member 94 and extend through a plurality of apertures therethrough to result in a row of guides 97 at the underside of the plenum chamber floor member 94. This arrangement is shown in Figures 8B and 8C. In Figure 8B, the sprung wire members 98 are shown attached to the top surface of the plenum chamber floor member 94. In the present example, the sprung wire members 98 are arranged in pairs, such that two guides 97 are provided by each spring clip 98. In Figure 8C, the guides 97 formed by the protruding portions of the sprung wire members 98 are shown at the underside of the plenum chamber floor member 94. Each guide 97 is advantageously positioned so as to interface with the groove 108 in the plastics material in the upper face 104 of a processing cartridge 43 as shown in Figures 5A-5C to aid correct alignment and to facilitate insertion of the processing cartridge during insertion of the cartridge into the shelf 41. The use of the spring clip as a guide 97 also serves to urge the processing cartridge downwards to provide a secure mounting of the processing cartridge 43, to take account of manufacturing and operational tolerances and to assist in insertion of the processing cartridge where an operator does not align this absolutely correctly.

A further row of cartridge guides 99 can be provided at the upper surface of the base 51 of the shelf 41. In the present example, as shown in Figure 8D, these guides 99 have a rail like form, which can be achieved by punching or stamping through the base 51 of the shelf 41. In this example each guide, or rail, 99 includes a pair of upstands separated by an aperture 100 through the base 51. The size of the aperture 100 can correspond to the width between the upstands. The separation of the

upstands is selected so that the overall width of the resulting rails is slightly less than the width of a groove formed in the lower face of an information processing cartridge 43. Thus, each guide 97 is advantageously arranged so as to interface with the groove 1113 in the plastics material in the lower face 104 of a processing cartridge 43 as shown in Figures 5A-5C to aid correct alignment and to facilitate insertion of the processing cartridge during insertion of the cartridge into the shelf 41.

In the present example, where the guides 97 and 99 are formed from metal, the provision of the grooves 108 and 113 in plastics material at the upper and lower faces, respectively, of each information processing cartridge 43 (see Figures 5A–5B) results in a combination of metal and plastics materials that can provide a low friction interaction, facilitating insertion of the information processing cartridges.

If, for example, the information processing cartridge enclosure is made of a metal, it may be undesirable to provide metal guides to avoid a metal to metal interaction. In such a case, for example, it may be desirable to form the guides from a plastics material having a low coefficient of friction, such as polytetrafluoroethene (PTFE) or polythene. Plastics rails could be attached to the underside of the plenum chamber floor member 94 and/or on the upper surface of the base 51 of the shelf 41. In such an example, grooves on the upper and lower faces of the information processing cartridges 43 could then be formed of metal or plastics and still result in a low friction arrangement.

A CSSP/PSU divider 96 can be provided to the rear of the midplane 171 and can extend horizontally to the rear face 59 of the shelf enclosure 47. The CSSPs 71, when inserted, are supported by the divider 96. To aid the correct insertion of the CSSPs 71, CSSP guide pins 178 are provided on the midplane 171 at positions adjacent connectors 177 on the midplane 171 for connection to the CSSPs 71.

Respective positions 88 and 89 can be formed in the front face 57 and the rear face 59 at which first and second indicator boards 183 and 184 supporting the indicator LEDs 69 can be located. These positions 88, 89 therefore include an aperture through the respective face of the shelf enclosure 47 such that indicator LEDs 69 mounted onto a circuit board attached to the inside of the shelf enclosure 47 may be viewed from outside the shelf enclosure.

There now follows are more detailed description of the midplane 171.

As mentioned above, the midplane 171 connects all the elements of a shelf together, including, in the present example, up to sixteen information processing cartridges 43, up to two CSSPs 71, two PSUs 81 and the two indicator boards 183 and 184. In the present example, due to its location within the shelf enclosure, the midplane 171 is not configured to be swappable. Accordingly, to maximize the system reliability, the midplane is configured to provide as a high level of reliability as possible. To this end, the midplane is advantageously configured without active devices and to include the minimum number of decoupling capacitors consistent with good design practice (ideally zero).

The midplane supports a number of paths for various power and signal lines to interconnect the FRUs.

In the present example, each information processing cartridge 43 has a high speed information signal connection (e.g., a Gigabit (Gb) Ethernet SERializer/DESerializer (SERDES) connection) to each of the CSSPs 71, each connection consisting of two pairs of differential signals. In a conventional manner therefore, the tracking of the paths for these signals is arranged to keep the pairs well balanced and on a single signal layer (i.e. without vias) to support such differential signals at high frequency.

In addition, in the present example, each information processing cartridge 43 has a serial console connection to the CSSP cartridge 71. Each connection consists of two TTL (Transistor-Transistor Logic) level signals that make a transmit and return (TX and RX) pair.

Also, each PSU 81 has a management signal connection (e.g., a serial I2C (Inter-IC Bus) connection) to the CSSP cartridge 71 to control power and monitor environmental parameters. The I2C bus comprises of two signals SCL and SDL (serial clock line and serial data line). In addition, an I2C address programming pin is provided for the PSUs 81.

Each information processing cartridge 43 and PSU 81 can signal to the CSSP cartridge 71 that it is inserted by pulling to ground (GND) a respective Inserted_L signal (i.e., an active low signal). These signals are fed to the CSSP cartridge 71 via the midplane 171.

Each PSU 81 has five 12 Volt output rails. The routing from each PSU 81 is arranged so that a fault in any single FRU cannot completely interrupt the power to any other.

As mentioned above, the midplane 171 is provided with appropriate connector arrangements for receiving the connectors on the FRUs.

In the present example, the information processing cartridge 43 connects to the midplane 171 through a 40 pin Single Connector Attachment (SCA-2) connector as defined by the Small Computer Systems Interface (SCSI) standard. Accordingly, the midplane carries corresponding connectors 175.

In the present example, each CSSP cartridge 71 connects to the midplane 171 through a two right-angle 20 pair connector (e.g., 2mm HM-Zd connectors available from Tyco Electronics). The corresponding connectors 177 on the midplane are straight male parts with a power connector. A guide pin arrangement is provided in addition to the connectors to prevent misaligned modules causing bent pins during insertion. The guide pin also provides a leading ground. The CSSP cartridge 71 also connects to the midplane 171 through a right-angled 125 way 5 row 2mm connector. The connector 177 on the midplane 171 includes a straight male part. A guide pin arrangement is provided in addition to the connectors to prevent misaligned modules causing bent pins during insertion.

In the present example, as mentioned above, each PSU 81 connects to the midplane 171 through an SSI-MPS specification connector. The contacts are configured 5P/24S/6P with sequenced signal (S) and power (P) pins. Where the connector on the PSU is a 1450230-1 R/A male header, solder tails connector, the mating connector 179 on the midplane can be a 1450540-2 vertical receptacle, press-fit connector.

In the present implementation, indicator boards 183 and 184 (see Figure 8A) are provided at the front and rear of the system and are configured as FRUs. In this example they hold three system-level indicator LEDs 69 and include a FRU identity (FRU-ID) programmable read-only memory (PROM) each. Three LEDs 69 are present on the indicator board. There can, for example, be a white locator LED that can be switched by the user for locating the system; a green power-on LED to indicate when the system is powered; and an amber service-required LED to indicate a fault or other condition requiring servicing. These LEDs can be driven by the CSSP 71.

In the present example, identification information (FRU ID) for the midplane 171 is held on an I2C electrically erasable programmable read only memory

(EEPROM) in the front indicator board 183. In addition to the I2C signals necessary to access the FRU ID EEPROM, the CSSPs 71 provide a current limited supply to the indicator boards 183 and 184 via the midplane. The indicator boards 183 and 184 are also provided with an I2C address programming pin. Depending on the implementation, FRU ID information can be stored instead, or in addition, on the rear indicator board 184.

As the FRU-ID for the midplane 171 is held on one or both of the indicator boards 183 and 184, the midplane can be a totally passive unit. The FRU-ID PROMs communicate with the CSSPs 71 via an I2C bus. Each device on the bus has a separate I2C address. The lower three I2C address bits of the EEPROMs used are available as pins on the device, to allow programming with resistors. The least significant bit of this address (A0) is passed to the midplane via the corresponding connector. This allows the midplane 171 to program the address of the FRU-ID differently for the front and rear indicator boards 183 and 184, by pulling the address low for the front board and high for the rear indicator board 183. This ensures that both EEPROMS are available on the bus, at different addresses. The FRU-ID for the midplane can be stored on either front or rear EEPROM, but the present example the FRU-ID is stored in the EEPROM on the front indicator board 183. The EEPROM can be 8kByte or larger.

As mentioned above, the midplane 171 includes openings 185 to provide a ventilation path for cooling air passing through the shelf 41. The cooling air passing through the shelf 41 via the midplane 171 can be driven by means of fans provided in each of the information processing cartridges 43 and the power supply modules 81. The openings 185 shown in Figures 8B, 9A, 9B and 9C form schematic representations of openings in the midplane 171. In practice, the openings could have any form (i.e., a series of large openings, or a number of small perforations), arranged on the midplane to align with corresponding openings or ventilation apertures in the

various field replaceable units 43, 71 and 81. In this way, the path of the airflow from the front of the shelf to the back of the shelf can be configured to be as efficient as possible, depending on the detail configuration of the fan units and the ventilation openings or apertures in the information processing, switch, service processor and power supply unit modules 43, 71 and 81. Providing the fan units in the field replaceable units 43, 71 and 81, contributes to the aim of maintaining the chassis 49 and the midplane 171 of the shelf 41 free of active components, thereby minimizing cost, and facilitating maintenance. Also, by providing the fan units in each of the field replaceable units, merely inserting and removing field replaceable units automatically adapts the flow of cooling air to the number and type of field replaceable units inserted in the shelf 41.

As described above, in the present example each of the FRUs is designed to be a non-user serviceable unit. Thus each FRU presents the user with a "sealed" unit which may be inserted into and removed from the shelf 41 as desired or required. If a FRU ceases to be operable, then the user has a choice only of returning the FRU to a supplier or service company for repair or of discarding the non-operable unit. As the FRUs are non-user serviceable, there is no requirement for a skilled technician to be employed in inserting or removing the FRUs into or from a shelf 41. Thus each FRU is designed such that a non-skilled person should have difficulty in causing damage to the FRU during handling. Moreover, the configuration and construction of the FRUs (e.g., the injector/ejector levers, the grooves in the enclosures of the information processing units, etc), of the shelf enclosure and the midplane (e.g., the guide rails to guide insertion of the FRUs, the locating pins, etc) contribute to facilitating easy insertion and removal of the FRUs.

Shown in Figure 10 is an example of the flow of cooling air through the shelf 41 and FRUs 43, 71, 81 mounted therein.

In this example, the cooling air passing through the shelf 41 is drawn generally in a front to rear direction through the shelf 41 by cooling fans mounted within the CSSPs 71 and the PSUs 81. Two separate flow paths for cooling air are provided in this example. The first, indicated as flow path $\alpha\alpha$ by dotted lines 77 provides cooling air to the CSSPs 71. The second path, indicated as flow path $\gamma\gamma$ by dotted lines 78 provides cooling air to the information processing cartridges 43 and PSUs 81.

The flow of cooling air along path $\alpha\alpha$ enters the shelf 41 through the aperture(s) 65 in the front face 57 of the shelf enclosure 47 into the plenum chamber 66. Further apertures (not shown) could also be provided in the side faces 53 and 55 of the shelf enclosure 47 to allow air to flow into the plenum chamber 66. This air then flows through the plenum chamber 66, and passes over the top edge of the midplane 171 to reach the perforations 133 of the front face of the CSSPs 71. The cooling air then passes through the CSSPs 71, providing cooling to the components thereof before passing out of the CSSPs 71 through the perforations 135 in the rear face of the CSSPs 71 thus being exhausted from the shelf 41. This flow of cooling air along flow path $\alpha\alpha$ is driven by fans 79 mounted within the CSSPs 71. In the present example, a pair of fans 79 is provided within each CSSP 71 and is mounted against the rear face thereof.

Air flowing along path $\alpha\alpha$ is impeded from flowing around the processing cartridges 43 by plenum chamber floor member 94 and is impeded from flowing to the PSUs 81 by CSSP/PSU divider 96. This flow path $\alpha\alpha$ therefore ensures that air flowing to the CSSPs 71 is not warmed by passage though the processing cartridges 43 and therefore provides maximum efficiency cooling to the CSSPs 71.

The flow of cooling air along path $\gamma\gamma$ enters the shelf 41 through the perforations 115 in the front face of the information processing cartridges 43. The air thus enters the information processing cartridges 43 and provides cooling to the

components thereof. Cooling fans (not shown) within the information processing cartridges 43 direct the cooling air to the processor (CPU) of the information processing cartridge and direct the flow of air in the cartridge thereby increasing cooling efficiency. The air then exits the information processing cartridges 43 through the perforations 117 in the rear face thereof. The air then passes through the apertures 185 through the midplane 171 to reach the PSUs 81. This cooling air then passes though the perforations 157 on the front and upper front faces of the PSUs 81 to enter the PSUs and provide cooling to components thereof. It will be appreciated from Figure 10 that the sloping rear of the upper face of the PSUs 81 increases the area over which air can be drawn into the PSUs, thereby reducing the back pressure on the air flowing through the shelf unit and aiding the cooling efficiency. The flow of cooling air along path $\gamma\gamma$ is driven by fans 85 mounted within the PSUs 81. In the present example, a pair of fans 85 is provided within each PSU 81 and are mounted against the rear face thereof.

Air reaching the PSUs 81 via path $\gamma\gamma$ will already have passed through the processing cartridges 43. Such air will therefore be already warmed above the ambient temperature outside of the shelf 41 by its passage through the processing cartridges 43. However, as the cooling requirement of the PSUs 81 is typically less than that for the CSSPs 71, this does not cause any difficulty in the operation of the PSUs 81, which are adequately cooled by this flow of pre-warmed air. The pre-warmed air passing through the apertures 185 through the midplane 171 is impeded from flowing into path $\alpha\alpha$ and entering the CSSPs 71 by the SCCP/PSU divider 96.

As will be appreciated by the skilled addressee, the arrangement shown in Figure 10 is illustrative only and other arrangements whereby, for example, a mixture of cool air from the plenum chamber 66 and warmed air from the processing cartridges 43 is provided to each rear-mounted FRU can readily be constructed.

With reference to Figure 11, there now follows a description of functional elements of an information processing cartridge 43 as contained within the information processing cartridge enclosure 101.

The information processing cartridge 43 includes a microprocessor 192 (a non-limiting example of a microprocessor that can be utilized in the present example is an UltraSPARC™ processor). The microprocessor is mounted on an information processing cartridge motherboard 191.

A configurable clock generator 193, configured as a programmable clock synthesizer employing a crystal, can be used to produce CPU clock signals, CLKA and CLKB. The clock frequency can be determined by jumper settings (not shown). A vectored interrupt controller (I-Chip) 194 and a configurable core voltage regulator module (VRM) 195 are provided.

In the present example, memory means for use by the processor 192 when executing instructions can be provided in the form of buffered dynamic random access memory (DRAM), for example configured as dual in line memory modules (DIMMs) 196 with a 72-bit data path with error correction codes (ECC), seated in two sockets on a riser card from the information processing cartridge motherboard 191. The memory capacity can be chosen to suit the processor addressable memory space. For example, in the present example, up to 4 Gigabytes (4GB) of addressable memory can be provided. Serial Presence Detect (SPD) auto-configuration is provided via a Service Management Bus (SMBus) over an I2C bus 197.

In the present example, a PCI bus architecture can be employed with a so-called SouthBridge bus bridge 199 with SuperIO and two Gb Ethernet Media Access Control (MAC) devices. As described above, however, other bus protocols (e.g., Infiniband) can be used. A 32bit PCI bus 198 can be provided from the

microprocessor 192. The SouthBridge 199 is a standard form of bus bridge, in the present example packaged in a 352 pin PBGA (Plastic Ball Grid Array) package, that provides the following functions: an SM Bus interface over the I2C bus 197 for access to the SPD (Serial Presence Detect) feature of the DIMMs that allows initialization of the memory controller; an Xbus interface for access via an Xbus 200 (which is a packet switched multiprocessor bus) to a PROM 201, a real time clock (RTC) 202 and an information processing cartridge service controller (hereinafter termed a Blade Service Controller (BSC)) 203; an IDE (Integrated Drive Electronics) interface that provides an ATA-100 (AT Attachment) IDE connection 204 to an IDE disk drive 205; and a serial console interface on a service bus 206 to the BSC 203 that is used for operating system functions including a console function with this embodiment.

For IO to the midplane 171, two AC-coupled Ethernet interfaces 207 and 208 are provided in the present example, which are packaged in a 316 pin PBGA. These Ethernet interfaces can provide a PCI attached Ethernet MAC capable of operation up to Gigabit Ethernet performance. The physical layer can be implemented using SERializer/DESerializers (SERDESs) 209 and 210. An example of a SERDES device is the TLK2201 transceiver manufactured by Texas Instruments, Inc. The SERDES devices use differential PECL TX+/- and RX+/- (Positive Emitter Coupled Logic Transmit and Receive) pairs to communicate to the switch portions of the CSSPs 71 over the midplane 171. The RX+/- pairs can be AC coupled at the information processing cartridge 43, the TX+/- pairs can be AC coupled at each CSSP 71. This facilitates hot-swap of the information processing cartridges 43 and the CSSPs 71.

Asynchronous serial connections 211 and 212 for communication between the BSC 203 and the Service Processor parts of the CSSPs 71 can be provided.

Internal data storage can be provided in the present example by a hard disk 205 with a capacity of 30GB or more rated for 24/7 continuous operation. The hard disk 205 is accessed using the primary IDE interface of the SouthBridge 199. The hard disk 205 can hold an operating system, for example a Solaris operating system, and other software and data for performing information processing using the main, or host, processor (CPU) within the information processing cartridge 43.

In the present implementation, the BSC 203 can be implemented as a microcontroller (e.g., a Hitachi H8 microcontroller). The BSC 203 can provide various functions, including for example: dual access (for the information processing cartridges and the CSSPs 71) to PROM 201 and EEPROM 213 for boot information and a FRU-ID for the information processing cartridge; channeling communication between an information processing cartridge 43 and the service processor part of the CSSPs 71; control of power on reset (POR), system reset and externally initiated reset (XIR) to the microprocessor 192; control of the power, service-required and ready-to-remove LEDs 69; upgrading of field-upgradable firmware, via the serial interface; a watchdog function for the operating system; monitoring the speed of a CPU fan 214; and communications with an EEPROM 215 and the operating system via the Xbus 200.

In the present example, the BSC 203 can be powered by a 5V service bus (SB) rail as soon as a CSSP 71 and a PSU 81 are fully inserted into the midplane 171, it then turns on other DC/DC converters to provide power to the remainder of the information processing cartridge 43. A BSC reset signal can be derived from a simple conventional power on reset (POR) generator that monitors a 5V supply rail.

In the present example a 1MByte Flash PROM 201 can be provided for storing boot variables for OpenBoot™ PROM (OBP) and Power-On-Self-Test (POST). Further OBP variables can be stored in a second 16kByte (16kB) I2C PROM 215,

accessible via the SouthBridge SM Bus port over the IC Bus 197. The PROM 215 can contain 8kByte for OBP variables and 8kByte of unused space. A 16kByte I2C EEPROM 213 that is accessible via the BSC 203 can contain BSC variables and FRU-ID variables. The EEPROM is nominally divided into 8kByte for FRU-ID and 8kByte for the BSC variables. Write protection for the FRU-ID is implemented by BSC firmware. Such write protection may be carried out by, for example, acknowledging instructions to write to the protected area, but not to carry out those write instructions.

An environmental monitor sensor 215 can be provided to monitor the CPU and ambient temperatures. This sensor can be accessible via the onboard I2C bus from the BSC 203.

The information processing cartridge 43 can be powered from two, diode commoned, 9V power supply rails 216 and 217. DC/DC converters 218 can be used to provide the voltage levels required by the information processing cartridge 43. The DC/DC converters 218 are supplied by dual 9V inputs 216, 217, individually fused 219, 220 and then diode commoned 221, 222. A 5V DC/DC converter can be turned on as soon as the FRU is fully inserted, with the BSC 203 and required portions of the SouthBridge 199 being powered (the 5VSB rail). A field effect transistor (FET) can be used to gate off the main 5V supply to the rest of the information processing cartridge 43. The DC/DC converter outputs and the main 5V FET can be arranged not to turn on until the BSC 203 turns them on via a signal from the SouthBridge 199. The SouthBridge 199 can be used so that if the BSC 203 is reset (by a watchdog timeout or after a firmware download) the state of the DC/DC converters 218 is not affected. When the remaining outputs from the DC/DC converters 218 are within specification, a PWR_GOOD signal can be asserted low to the BSC 203.

A SouthBridge resume circuit can be operable to run from 3V3, and a simple Zener diode dropper circuit can be used to generate 3V3 from the 5VSB supply.

When the FRU is inserted the inrush current can be limited, for example to <1A, and the rate of rise can be configured not to exceed a predetermined value (e.g., 20A/s) to provide a so-called soft start to facilitate hot-insertion. The intent is to prevent damage to the connectors and to avoid generating noise. A soft start controller 223, which controls a ramping-up of voltage levels, can be enabled when the predetermined signal (Inserted_L signal) is asserted low, this signal is on a short pin in the connector and is connected to ground (GND – not shown) through the midplane 171.

In the present example, a processor impingement fan (processor fan) 214 is configured to run at full speed to cool the information processing cartridge 43 and the fan. The speed of the processor fan and sink can be monitored by the BSC 203, using a tachometer sense pin on the microcontroller. In the event of the fan speed falling below a predetermined speed, or percentage of its nominal speed (e.g., 80%), the BSC 203 can be arranged to issue an alert. The nominal speed of the fan can be recorded as part of the BSC EEPROM contents.

The midplane connector 120 for the information processing cartridge 43 is used to establish the connection between the information processing cartridge 43 and the midplane. In the present example it supports up to 84 connections (pins) that will deliver SERDES outputs 224, 225, I2C signals 226, 227, and power 216, 217. Signal connections may be made through a right-angled connector. Power connections may be made through the information processing cartridge right-angled connector. The connector can be configured to facilitate hotswapping of the information processing cartridge, for example by having a low insertion force and/or guide pins to increase the ease of serviceability and prevent module misalignment during insertion.

Interrupts to the processor 192 can be encoded using an encoded interrupt vector mechanism. An I-Chip Emulator (ICE) 228 functions as an interrupt concentrator, receiving all system interrupts and encoding them as an interrupt vector according to an interrupt vector code utilizable by the processor 192. In the present example, where an UltraSPARC™ processor is used, the interrupt vector encoding may be based on a 6-bit interrupt vector code.

With reference to Figure 12, there now follows a description of an example of a combined switch and service processor (CSSP) 71. In the present example, each CSSP 71 provides the functionality of a Switch 73 and of a Shelf Service Processor, or Shelf Service Processor (SSP) 74.

Figure 12 provides an overview of the functional components of the CSSP 71 including functional components of the Switch 73 and functional components of the SSP 74. In the present example, most of the components relating to the Switch 73 are mounted on a Switch PCB 231, and the components relating to the SSP 75 are provided on a SSP PCB 232. However, it should be noted that the components located in the lower portion of the switch PCB 321 (i.e., that portion below the SSP PCB 232 as illustrated in Figure 12 logically belong to the SSP 74, rather than to the switch 73. It will be appreciated that such component arrangements are not compulsory for successful operation and that any other component arrangement over any number of component boards can be easily achieved using conventional component arrangement techniques.

Firstly, with reference to Figure 12, there follows a description of functional elements of the Switch portions 73 of a CSSP 71 as contained within the CSSP enclosure 121.

The midplane connector 141 on the CSSP 71 establishes the connection between the CSSP 71 and the midplane 171. In the present example, it supports up to 84 connections (pins) that will deliver SERDES outputs 265-268, I2C signals 310, 320, 321 and 322, and power 278, 279. Signal connections may be made through two 20-pair right-angled connectors. Power connections may be made through a right-angled connector. The connector can be configured to facilitate hotswapping of the board, for example with a low insertion force. The connector also uses guide pins to increase the ease of serviceability and prevent module misalignment during insertion.

A switch microprocessor 240 is provided, in the present example the microprocessor used is a Power PC (MPC8245) packaged in a 352pin Tape Ball Grid Array (TBGA) package. This microprocessor 240 supports between 1MB and 2GB of address space in the present example. It further includes an Embedded Programmable Interrupt Controller (EPIC) that provides 5 hardware interrupts (IRQs) or 16 serial interrupts. There are 4 programmable timers with cascade mode function. DRAM memory for the processor can provided in the present example by a commodity DIMM 242. The processor 240 can be connected to a 32bit PCI bus 241, which operates at, for example, 33MHz/66MHz.

A clock input to the processor 240 can be provided by a clock generator (CLK) 243. The CLK 243 can include a configurable clock generator (not shown) implemented as a programmable clock synthesizer employing a crystal used to produce CPU clock signals. The clock frequency can be determined by jumper settings (not shown). A vectored interrupt controller (I-Chip) (not shown) and a configurable core voltage regulator module (VRM) (not shown) can be provided that operate substantially as described above with reference to the like components of Figure 11.

In the present embodiment two switch ASICs (application specific integrated circuits) 244, 245 are provided (in the present example, BCM5632 Gigabit switch ASICs). Each ASIC can provide twelve GMII Interfaces (1 Gigabit Ethernet) (for uplinks and downlinks) and one 10Gb XGMII interface for chip-to-chip communication (bridging) 246 between the ASICs 244 and 245. Sixteen GMII 1Gb 'downlinks', in the form of serialized Gb Ethernet data, are provided through four quad SERDES 248-251 to allow each information processing cartridge 43 to communicate with the switch 73. Eight GMII 1Gb 'uplinks' are provided for external communication through two quad PHYs 253 and 254 (in the present example BCM5404 ASICs) and RJ45 connectors on the rear panel 122. The ASICs 244 and 245 are configured via a PCI interface (32bit/33MHz) to the PCI bus 241.

A Flash PROM 256 can store a real time operating system, and management and configuration data for the microprocessor. The Flash PROM 256 in the present example can be operable to hold 8MB - 16MB of data, depending on the software required. The flash PROM 256 can be operated via an on-chip XBus 258.

Also connected to communicate with the processor 240 via the XBus 258, a Real Time Clock (RTC) 259 can be provided for real-time functions with a back-up battery.

Also connected to the XBus 258 can be a UART (Universal Asynchronous Receiver Transmitter) 260 which in turn connects to a serial bus 261 for providing an asynchronous console connection from the switch 73 to the SSP 74 which can be accessed by the SSP.

An integrated MAC/PHY (Media Access Control/Physical) switch 271 can provides its own interface to the PCI bus 241. This MAC/PHY switch 271 can connects to a 10/100 Ethernet hub 272. The hub 272 can be operable to provide a

management interface to the SSP 74 and a connection from an external management network to the switch 73 and SSP 74 of a given CSSP 71. The connection from the integrated MAC/PHY device 271 to the SSP 74 can be coupled capacitively. A loopback mode can be provided by the MAC/PHY device 271 for system diagnostics. The hub 272 can connect to an RJ45 connector 273 on the rear panel 122 of the CSSP enclosure 121.

An 8kByte I2C EEPROM 262 can be used to store the FRU-ID and is accessible by the SSP portion 74 of each CSSP 71 via a serial bus 263 and the midplane 171. The upper 2kByte of the EEPROM 262 can be configured to be write protected.

An I2C Redundant Control Register (RCR) 275 can be used to provide an alternate, redundant path for powering-down the CSSP 71 and Shelf Level Indicators 69 mounted on the front 57 and rear 59 panels of the shelf 41. The I2C RCR 275 can be accessible by both the SSP 74 of the CSSP 71 containing the RCR and the SSP 74 of a further CSSP 71 connected via the midplane 171 via an I2C bus 276. In the present example, a device suitable for use as the RCR 275 is a Phillips PCF8574 IC.

With continued reference to Figure 12, there now follows a description of functional elements of the Shelf Service Processor (SSP) portion 74 of a CSSP 71 as contained within the CSSP enclosure 121 and provided on an SSP PCB 232.

In the present example, communication between the Switch PCB 231 and the SSP PCB 232 is facilitated by an interboard connector pair 298 and 299. It supports connections (pins) for I2C signals, 10/100 MAC/PHY output, and power. As described above, the switch PCB 231 carries the components associated with the switch, and it also carries the power, FRU-ID and environmental monitoring components along with the connectors for connections to the midplane 171 and

external connectors. Thus, in the present example, all SSP components requiring a connection to the midplane 171 or an external connection have signal paths routed through the connector pair 298, 299 and via the switch PCB 231 to the relevant midplane or external connectors.

In the present example, the SSP 74 includes a microprocessor 301 (e.g., a Power PC (MPC8245) processor) mounted on the SSP printed circuit board (PCB) 232. The processor 301 can be connected to a PCI bus 302, the present instance a 32 bit bus that operates, for example, at 33MHz/66MHz.

A clock input to the processor 301 can be provided by a clock generator (CLK) 303. The CLK 303 can comprise a configurable clock generator (not shown) implemented as a programmable clock synthesizer employing a crystal used to produce CPU clock signals. The clock frequency can be determined by jumper settings (not shown). A vectored interrupt controller (I-Chip) (not shown) and a configurable core voltage regulator module (VRM) (not shown) can be provided that operate substantially as described above with reference to the like components of Figure 11.

The processor 301 can be provided with a DRAM memory 305. The memory capacity can be chosen to suit the processor addressable memory space. In the present example, 8 MB of DRAM memory is provided.

An integrated MAC/PHY switch 306 can provide its own interface to the PCI bus 302. The MAC/PHY switch 271 can be connected to 10/100 Ethernet hub 272 via the interboard connectors 298, 299. A loopback mode can be provided by the MAC/PHY switch 306 for system diagnostics.

Octal UARTs 308 and 309 can be connected between the PCI bus 302 and the interboard connector pair 298, 299. The signal path can be continued from the interboard connector pair 298, 299 to serial connections 310 on the midplane connector 141 on switch PCB 231. The Octal UARTS 308, 309 can facilitate serial communications between the SSP 74 and each of the processing cartridges 43.

Also connected to the PCI Bus 302 can be a dual UART (DUART) 312 that in turn can connect via the interboard connectors 298, 299 to serial bus 261 for providing an asynchronous console connection from the SSP 74 to the switch 73. The DUART 312 can also have an I2C connection to an external connector on the rear face 122 of the CSSP enclosure 121. The external connector can provide a common operating system/boot console and command port 311.

Connected to the processor 301 via an XBus 314 can be a Flash PROM 315. The Flash PROM 315 can store a real time operating system, and management and configuration data for the microprocessor 301. The Flash PROM 315 can be operable in the present example to hold up to 2MB of data, depending on the software required.

Also connected to the processor 301 via the XBus 214 can be a real time clock (RTC) 316 for real-time functions with a backup battery. The RTC 316 can also provide 8kByte of non-volatile random access memory (NVRAM), in the present instance implemented as an EEPROM. This can be used to contain information such as the FRU-ID, a serial number and other FRU information.

To facilitate I2C communications between the SSP 74 and the other CSSP 71, the midplane 171 and the PSUs 81, a multiplexer 318 can be provided. The multiplexer 318 can have a single I2C connection to the processor 301 and connections, via the interboard connector pair 298, 299 and the midplane connector 141 to both PSUs 81, the midplane 171 and the other CSSP 71.

The processor 301 can also comprise an embedded DUART to provide a redundant serial link to the SSP 74 of the other CSSP 71. Although it would be possible to implement this link using an external DUART, the advantage of using an embedded DUART is that the connection to the other CSSP is reliable and therefore likely to be functional. Where the embedded DUART link does not use the I2C Multiplexer for communications to the other CSSP, a common mode of failure for both the SSP – SSP I2C links can be avoided, it being assumed that the processor 301 is likely to be functional even if both embedded DUART channels are non-functional.

The CSSP 71 can powered from two, diode commoned, 9V power supply rails 278 and 279. DC/DC converters 281 can be used to provide the voltage levels required by the CSSP 71. The DC/DC converters 281 can be supplied by dual 9V inputs 278, 279, individually fused 285, 286 and then diode commoned 287, 288. A soft start controller 283 can be provided to facilitate hot-insertion. A 5V DC/DC converter (I2C power regulator) 282 can be turned on as soon as the CSSP 71 is fully inserted. A 3.3V DC/DC converter can be turned on when instructed, for example through SSP service software, by asserting low an appropriate signal (ON_L – not shown). The 3.3V converter can be arranged to turn on a converted for 2.5V, 1.2V, and a processor core voltage rail (Vcore) when the voltages are within an appropriate range.

When the CSSP 71 is inserted the inrush current can be limited, for example to <1A, and the rate of rise can be configured not to exceed a predetermined value (e.g., 20A/s) to provide a so-called soft start to facilitate hot-insertion. The intent is to prevent damage to the connectors and to avoid generating noise. A soft start controller 283, which controls a ramping-up of voltage levels, can be enabled when the predetermined signal (Inserted_L signal) is asserted low, this signal is on a short pin in the connector and is connected to ground (GND – not shown) through the

midplane 171 until one of the supplies is removed. These circuits can be configured to withstand an overvoltage at their inputs whilst the input they are feeding is not powered, without any leakage to the unpowered circuit. A sense circuit can detect if the voltage has dropped below a threshold, for example 2.0V, as a result of a blown fuse, a power rail going down, etc. The DC/DC converters 281 can be protected against short circuit of their outputs so that no damage occurs.

The I2C regulator 282 can be powered as soon as the CSSP 71 is fully inserted into the midplane 171. This can be facilitated through short pins connected to the soft start controller 283, which controls a ramping-up of voltage levels. The other DC/DC regulators can be turned on, for example by SSP software.

A pair of fans 290, 291 can provide cooling to the CSSP 71. The fans 290, 291 can be configured to run at full speed to prevent overtemperature conditions by minimizing the temperature of the internal components and the fan. The speed of the fans 290, 291 can be monitored by the SSP 74 through an environmental monitor 295 on the switch board 231. The environmental monitor 295 can be alerted in the event of the fan speed falling below a predetermined value (e.g., 80% of its nominal speed). The fan can provide tachometer outputs to facilitate the measurement of fan speed.

LED indicators 137 can be provided, for example with a green power LED, an amber LED for indicating that service is required and a blue LED for indicating that the switch is ready to be removed. LED indicators integrated on 2x4 stacked RJ45 connectors on the rear face of the CSSP 71 can be arranged, for example, to show green continually when the link is present and flash green when the link is active.

The environmental monitor ENV MON 295 can be provided to maintain operational integrity of the CSSP 71. The ENV MON 295 can include limit values in limit registers and can monitor, for example, temperature within the CSSP enclosure

121, the CSSP power rails, including the 12V, 3V3, Switch Processor Core Voltage, CSSP Processor Core Voltage and the two 9V power feed rails 278, 279 from the midplane 171. The outputs of the DC/DC converters 281 can be fed in to A/D inputs of the ENV MON 295 for Watchdog comparisons to be made to the voltage limits set in the limit registers. As noted above, the ENV MON 295 can also monitor the operating speeds of the fans 290 and 291. The ENV MON 295 can communicate with the SSP 74 of both CSSPs via an I2C bus 296.

For IO to the midplane 171 shown in Figures 8A-10, the midplane connector 141 can include sixteen 1Gb Ethernet connections 265-268 from four quad SERDES 248-251 and the I2C bus lines 596.

The SSP 74 can access the I2C devices (FRU-ID EEPROM, 8-bit I/O expansion chip, and the system hardware monitor) through the midplane 171.

For external IO, rear panel Gb Ethernet connections can be provided from the two quad PHYs 253, 254 to 2x4 stacked RJ45 connectors 139 (to give 8 uplinks). Each port can be an independent 10/100/1000 BASE-T (auto negotiating) port. The PHY devices 253, 254 can operate in GMII mode to receive signals from the 8-Gigabit interfaces on the ASICs 244, 245.

The Power Supply Units (PSUs) 81 can configured such that when two or more PSUs 81 are connected in parallel in the shelf 41, failure of any one of the paralleled units shall not affect system operation. Moreover, one of the PSUs can be installed or removed from a "live" system with or without input power applied. The outputs can have overcurrent protection.

The PSU can have an I2C interface to provide power supply status via the midplane 171. The PSU can have an internal temperature sensor that reports via the

I2C interface. The PSU fan speed can also be monitored and errors are reported via the I2C interface. Overvoltage and overcurrent sensors can also report via the I2C interface.

There now follows a description of aspects of an example of a power supply 81 with particular reference to Figure 13.

When a power supply (e.g. mains power, or UPS type protected power) is connected to the cable connector 83, transformer, regulator and rectifier circuitry 400 can operate to generate a DC output (in the present example, 12V DC) from the input (in the present example 230/240V 50Hz AC or 110V 60Hz AC).

In order to ensure good cooling reliability within the shelf 41, each PSU 81 can have a pair of cooling fans 402, 403 located at the rear of the PSU enclosure as described above with reference to Figure 7. As a failsafe measure against a failure of one PSU 81, the fans of each PSU 81 can be powered by both PSUs 81. Thus, in the event that one PSU 81 is non-operative to produce the DC supply for the shelf 41, as long as the other PSU 81 is operative, not only will all components of the shelf 41 be powered as normal, the fans of both PSUs 81 can continue to run.

As shown in Figure 13, this dual powering of cooling fans 402, 403 can be effected by providing a power supply line 404 from the transformer, regulator and rectifier circuitry 400 to power both fans 402, 403. Also, first and second separate power lines 410, 412 from the other PSU 81 can provide duplicate power supply to the first and second fans 402, 403 respectively. The fan 402 can thus be powered by a diode commoned supply from line 404 and a diode commoned supply from line 410. Diode protection can be provided by diodes 405 and 411 respectively. The speed of the fan 402 can be controlled by a speed controller 408. Similarly the fan 403 can be powered by a diode commoned supply from line 404 and a diode commoned supply

from line 412. Diode protection can be provided by diodes 406 and 414 respectively. The speed of the fan 403 can be controlled by a speed controller 409.

The two speed controllers 408, 409 can in turn be controlled by a data input from each CSSP 71 received via an I2C bus connection (not shown in Figure 13). Power supply lines carrying DC power for the other FRUs of the shelf 41 are shown in Figure 13 as power line 416. All power connections to and from the PSU 81 can connect to the midplane 171 when the PSU is inserted in the shelf 41 via the midplane connector 163. In the present example the PSU 81 connects to the shelf through a 5P/24S/6P configuration SSI-MPS compliant right angle connector 163 at the front face 147 of the PSU 81. Connectors for the I2C interface can also be provided.

The input power line 410 and 412 for each fan 402 and 403 can be provided with a softstart module 4131 and 4132 respectively, to allow for hot insertion of the PSU 81 into the shelf 41. The softstart modules 4131 and 4132 can be controlled, for example, by pulling a signal to ground (e.g., a "mated" input line 4151 and 4152).

Where the two input power lines 410 and 412 are separate lines having separate softstart provision, there is no common failure mode for the backup method of powering the fans 402, 403. Thus even if the PSU 81 ceases to be operable to generate the DC supply, and a component (power line or softstart module for example) fails in the supply path from the second PSU 81 to the fans 402, 403, at least one of the fans 402, 403 can remain operational as the shelf 41 still receives the cooling effect of three PSU fans.

In the present example, the power supply has four rear panel LED indicators 137. A blue "Ready to Remove" LED can be driven by the I2C interface and indicate that the power supply may be removed from the system. An amber "Service Required" LED can be driven by the I2C interface and indicate that the power supply

is in a fault condition: any output out of range, over-temperature or shutdown. A green "DC Output-OK" indicator can be driven by internal power supply circuits and show that the main 12 volt supply is functioning. The LEDs can remain lighted when individual outputs are in the current limited mode of operation. A green "AC Input-OK" indicator can be driven by internal power supply circuits and show that AC input power is within normal operating range.

With reference to Figure 14, there will now be described an example of data connectivity between the FRUs and midplane of the shelf 41. Power transmission paths are not illustrated in Figure 14. However, it will be appreciated that to facilitate the maximum component redundancy of the shelf 41, each PSU 81 can independently provide power to each FRU.

In the present example each of the processing cartridges (blades) 43 connects to the midplane 171 via a pair of information signal connections (e.g. Gb Ethernet links) 224, 225 and a pair of serial management signal connections 226, 227. Connections within the midplane 171 can ensure that each Ethernet link 224 is directed to a connection 265-268 from the midplane 171 to a first switch 73, and that each Ethernet link 225 is directed to a connection 265-268 from the midplane 171 to a second switch 73. Thus one Ethernet link can be established between each processing cartridge 43 and the switch 73 of each CSSP 71. Further connections within the midplane 171 can ensure that each serial connection 226 is directed to a connection 310 from the midplane 171 to the first SSP 74 and that each serial connection 227 is directed to the second SSP 74. Thus one serial link can be established between each processing cartridge 43 and the SSP 74 of each CSSP 71. As mentioned earlier, information signal connections other than Gb Ethernet connections (e.g. Infiniband connections) could be employed in other examples.

A plurality of serial connections can connect each SSP 74 to the other. Serial lines 320, 321 can connect each SSP 74 to the midplane 171 and connections within the midplane can connect the two sets of lines together. To provide a control interface from the SSPs 74 to the PSUs 81, serial lines 322 can connect each SSP 74 to the midplane 171 and connections within the midplane 171 can connect to serial lines 324 from the midplane 171 to each PSU 81.

An example of data and control connectivity of the shelf 41 to and from computer systems external to the shelf 41 when the shelf 41 is arranged for use within a multiprocessor server system such as that described above with reference to Figure 1 will be described with reference to Figure 15.

As summarized above with reference to Figure 14, in the present example each processing cartridge, or blade, 43 is connected to the switch 73 of each CSSP 71 by an information signal connection (e.g. a 1Gb Ethernet link) formed by a combination of links 224, 225 from the processing cartridge 43 to the midplane 171, connections within the midplane 171 and links 265-268 from the midplane 171 to each switch 73.

Further, in this example a set of serial management signal connections comprising links 320, 321 and connections within the midplane 171 connect the SSP 74 of each CSSP 71 to the SSP 74 of the other CSSP 71.

To provide external data connectivity between the shelf 41 and an external core data network 330, in association with which all information processing performed by the processing cartridges 43 of the shelf 41 is undertaken, connections 331 can be formed between the core data network 330 and the eight 1Gb Ethernet ports 139 provided on the rear panel 122 of the CSSP enclosure 121.

In the present example, the connections by means of which control and configuration of the shelf 41 are performed are entirely separate to the connections to the core data network 330. Therefore, a first external switch 335 can connect to a management (I2C) port 273 of the first CSSP 71 and a second external switch 336 can connect to a management (I2C) port 273 of the second CSSP 72. As described above with reference to Figure 12, the management port 273 can provide a management network interface to both the switch 73 and SSP 74 of each CSSP 71. The external switches 335, 336 can each be connected to each of a pair of System Management Server (SMSs) 338, 339. The SMS is not essential to the operation of the shelf 41, but use thereof aids optimal operation of the shelf 41. In a typical multiprocessor server system a plurality of shelves 41 may be connected together via the core data network 330 under the control of a single management network utilizing one set of SMSs 338, 339. A set of SMSs 338, 339 may comprise a single SMS (as well as a plurality thereof). However use of at least two SMSs enables redundancy of components, therefore increasing overall system reliability.

A serial interface control 343 operable under telnet protocol control is also connected to the shelf 41 in the present example. This can provide a common operating system/boot console connection to the SSP 74 of both CSSPs 71 via the RJ45 connector 311 on the rear panel 122 of each CSSP enclosure 121.

It will be appreciated from the above that a flexible and scalable modular computer architecture has been described. In the described example up to 16 information processing cartridges, or blades 43, can be configured as sealed FRUs on a single shelf 41, the number of blades being chosen according to customer requirements. Each blade has its own processor and random access memory. If, for example, there is a maximum of 2Gbytes of memory per information processing cartridge, and one processor per blade, 16 processors (16P) with 5.33 processors per unit height (1U) and a total of 32GB of memory per shelf can be provided.

In the present example, the shelf 41 incorporates redundant combined switch and shelf service processor modules (CSSPs) 71 and redundant power supply units (PSUs) 81 separate from the blades 43. As the power supplies are carried by the shelf, the information processing cartridges can be kept compact and inexpensive. Also, as a result, they can be powered by DC power only, via the midplane 171.

Also, as mentioned earlier, the FRUs (e.g., the information processing cartridges, or blades, 43, the CSSPs 71 and the PSUs 81) can all be configured as sealed units that do not contain any internal FRUs themselves and do not contain user serviceable items. The enclosures of the FRUs can be arranged to enclose all of the functional components of the FRU with only electrical connectors being externally accessible and with indicator LEDs being externally visible as well.

These factors can all contribute to keeping the cost of the FRUs low, as well as that of the overall system. The modular approach with the use of sealed modular field replaceable units for providing system functionality and with non-field replaceable units designed with a minimum possible number of active components enhances reliability. Moreover, easy and rapid maintenance is facilitated in the event of a failure of a FRU by simple replacement of that FRU, further reducing the cost of ownership.

Thus, it will be appreciated from the above description that the provision of a rack mountable shelf, that includes power supplies, a shelf service processor and switches in modular units, for carrying a number of processing cartridges, wherein the number of processing cartridges can be chosen according to customer requirements, provides a flexible and scalable computer configuration. The balancing of the load between the processors of the processing cartridges can be effected by software using conventional principles.

A configuration as described provides an easily scalable processor architecture, whereby the processing power provided by a complete system based on the information processing cartridge/information processing cartridge carrier architecture can be scalable from moderate to very high capacity through the simple addition of further information processing cartridges.

As illustrated in Figure 16, an example of the external connections from a shelf 41 can be in the form of two active information signal connections (e.g., Ethernet connections) 350 and 351, two active power connections 353 and an active/standby pair of management connections 354. With regard to the management connections, each connection comprises a serial connection and a network (e.g. Ethernet or Infiniband) connection. It is possible to connect to either the active or the standby connection, as the incoming signal will be internally routed to whichever management controller (CSSP) is the current master. It will be appreciated, therefore, that the connections to a shelf can be kept to a minimum. It will further be appreciated from the configuration shown in Figure 16 that the system is scalable beyond a single shelf unit 41.

Figure 17 illustrates how a plurality of shelves can be configured within one (or more) racks to provide even higher processing power. Such a constellation of shelves to provide a large grouping of servers is sometimes termed a "web farm" or "server farm" 360. As shown in Figure 17, the web farm comprises a plurality of shelves 41 that each carry a plurality of blades 43. Also provided are a plurality of Network Attached Storage devices (NAS) 373 for providing storage for critical data, e.g. email data storage, for the web farm. The NASs 373 are not required if there is no critical data to be stored, e.g. if the web farm is operating solely to provide web caching services.

Management control of the web farm 360 can be provided through a pair of System Management Servers (SMSs) 362. Each SMS 362 can be connected to a management network via a link 366 and to a management console 365. The SMSs 362 can communicate with the individual shelves 41 via a pair of management switches 364. Each shelf 41 and NAS 373 can be connected to each management switch 364 via a connection 367. Thus dual redundant management connections can be provided to each shelf 41 and NAS 373.

Flow of data to and from the web farm 360 can be provided through a pair of data switches 369. Each data switch 369 can be connected to a consumer network via a link 370. It is to be understood that the consumer network can be a larger data network to which the web farm 360 is connected. This network can be an office or corporation intranet, a local area network (LAN), a wide area network (WAN), the Internet or any other network. Connections between the data switches and the shelves 41 can be facilitated by connections 371. It is to be noted that as each shelf has its own switching capability, there is no need for each shelf 41 to be directly connected to the data switches 369. Connections can also be provided to connect the NAS units 373 to the shelves 41. The topology used for interconnection of the data switches 369, shelves 41 and NASs 373 can be any topology providing at least one connection of any length between every possible pair of units. Complex topologies arranged to minimize the maximum connection length between any two given units in the web farm can be used.

The web farm 360 comprising a plurality of shelves 41 with or without a plurality of NASs 373 can suitably be used as any or all of the entry edge server group 9, web edge server group 15 and application servers 19 described above with reference to Figure 1.

As an alternative to providing critical data storage within a NAS 373, such storage can be provided within one or more NAS cartridges fitted into one or more of the shelves 41 in place of processing cartridges 43. Another alternative is to provide a server shelf with local storage (such as a RAID array (Redundant Array of Inexpensive Disks) in place of the NAS 373.

Thus there has now been described an example of a fully configurable computing system based on a plurality of self contained field replaceable units (FRUs) and scalable from a single processing cartridge with power supply and switching capability to a multiply redundant multiprocessor server system with full system management capability extending over a number of co-operably connected server shelves. It will of course be readily apparent to the skilled reader that many of the specific features specified in the above description are in no way limiting and a variety of alternatives may be produced using only ordinary skill and common general knowledge. Non-limiting examples of example modifications which may be made to the above described system are discussed hereafter.

There is no limit placed on the processing cartridges as to what software they should run. Each module within a shelf or farm may run under the same operating system, or a plurality of different operating systems may be used. Examples of possible operating systems include Sun Microsystems' Solaris® OS or another UNIX™-Type OS such as Linux™, MINIX™, or Irix™, or UNIX™ or a Microsoft OS such as Windows NT™, Windows 2000™, Windows ME/98/95™, Windows XP™.

It is also not necessary that each processing cartridge within a shelf or farm be configured to run the same program software. For example, individual processing cartridges may be configured to execute, for example, fileserver software, mailserver

software, webhosting software, database software, firewall software, or verification software.

Although it has been described above with reference to Figures 4, 8, 12, 13 and 14, that functionality of a switch and of a shelf service processor should be provided within a single combined switch and service processor unit, this is not essential and separate switch and shelf service processor field replaceable units may be used.

Although it has been described above that a pair of PSUs and a pair of CSSPs may be provided so as to enable dual-redundancy, further PSUs and CSSPs may be provided so as to increase FRU redundancy further, thus providing statistically higher reliability.

In the power supply circuitry in each of the blades and CSSPs, two voltage sense circuits may be provided after the fuses and before the diodes, to prevent a latent fault caused by a failed fuse going undetected until one of the PSUs is removed or taken offline. Such circuits may configured to withstand an overvoltage at their inputs whilst the input they are feeding is not powered, without any leakage to the unpowered circuit.

Although it has been described above with particular reference to Figure 11 that the processing module may be based on an UltraSPARC™ processor, this is not limiting and any other processor having sufficient processing capacity to undertake the tasks required of a particular processing cartridge may be used. Alternative processors include, but are not limited to, Intel x86 series and compatible processors, AMD x86 compatible processors, Alpha processors and PowerPC processors. The particular example of an x86 compatible processor is described in more detail with reference to Figure 18. In Figure 18, the parts corresponding to those of the UltraSPARC™ based

system of Figure 11 have the same reference numerals and will not be described again here. In a system based on an x86 compatible processor, the processor 378 itself communicates with the other components, including the memory 196 and PCI bus 198 via a Northbridge 379. The Northbridge 379 also includes an interrupt controller, so no separate interrupt concentrator is required. The other components of the processing cartridge could be substantially the same as for the UltraSPARC™ based system described above.

Although it has been described above that each information processing cartridge comprises a single microprocessor, this is not a limiting case as each or any of the information processing cartridges may have more than one microprocessor arranged to share common storage resources to operate synchronously (in lockstep) or asynchronously. Also, it is not necessary that all information processing cartridges inserted into a shelf at a given time are identical, rather a variety of different blade architectures may be used simultaneously.

The provision of the functions of both Switch and Shelf Service Processor within a single FRU in the present example provides a facility within a single shelf 41 for dual redundancy in both functions in fewer different FRUs. As will be appreciated, there is no restriction that these two functions must be provided within a single FRU and division of the two functions into separate FRUs would present no difficulty to the skilled addressee.

The backup power provision for the cooling fans of the PSUs is described above with reference to Figure 13. Although it is described that the backup power supply to each fan should be independent of the other, if the risk of common mode failure of backup power supply is judged to be low, or is of low importance, then the backup supply to each fan may be provided in common with the backup supply to all other fans.

Also, an information processing module forming a field replaceable server blade can include a processor and memory can be configured by means of software, firmware or hardware to provide a special purpose function. By way of examples only, an information processing module can be configured to perform the function of one or more of a firewall, or a load balancer, encryption and/or decryption processing, an interface to a secure network, e.g. a virtual private network (VPN), a specialized switch with wide area network (WAN) connectability.

Also, a storage blade may be provided. The storage blade can be configured to be mountable in a server blade receiving location in a blade server carrier. The storage blade can comprise storage blade connectors configured for interconnecting with carrier connectors on the server blade carrier, whereby the storage blade is interchangeable with a server blade. A carrier, or shelf, for a server system can be arranged with a plurality of blade receiving locations for receiving blades, wherein the blades can be storage blades or information processing blades. The server system can be self configuring on receipt of the blades according to the type of blade received in each said location. To achieve this the blade service controller in each blade can be operable to communicate with a shelf service processor to perform said configuring.

During normal operation of a server blade 43, the operating system of the main processor domain of the server blade (for example, the operating system for the processor 192) is operable to transmit heartbeat operations at predetermined intervals to the service processor (SSP) 74 in the CSSP 71. The SSP 74 is operable to monitor the heartbeat signals and treat these as "pat" signals for patting a watchdog timer implemented on the processor 301. In the present instance, the watchdog timer can be implemented in software on the processor 301, as the SSP is a separate system from the server blade(s) being monitored. Alternatively, a hardware timer could be provided in the SSP. In the event that a heartbeat signal is not provided within a

predetermined interval, as represented by an overflow timing of the watchdog timer, the SSP 74 will note this as being a failure of the server blade and will take some remedial action. The remedial action could be to attempt to restart the server blade by forcing the server blade to abort and to restart, or by turning off the server blade. In either case, an appropriate alert can be recorded and/or sent to the management console 365 to indicate that an error in the server blade concerned has occurred.

The monitoring of the system module can be important to maintain overall system reliability, especially but exclusively in a system such as that described above that is horizontally scalable and wherein a large number of a system modules can be operable. Such a watchdog timer can operate during the normal operation of the operating system on the server blade. However, as mentioned in the introduction, such an approach to monitoring the health of a blade server cannot work when the operating system of the server blade is not operational.

For example, in a system such as that described above, an approach to dealing with a transient fault in one of the system modules (for example one of the blade servers 73) can be to restart the system module concerned. In a highly horizontally scaled system, it may be necessary to restart individual system units from time to time to deal with such transient faults.

During the state transitioning for such a restart, there will be a period when the system module operating system is not operational. The conventional watchdog approach is not able to detect if a system module were to hang during such state transitioning. Accordingly, there is a need to be able to detect a failed transitioning of a system module, for example a failed restart of a system module in order to enhancing overall system reliability and resource availability.

In the following, embodiments of the invention will be described which enable system health monitoring to be maintained during state transitioning of the server blade, that is during changes in states other than the normal operating state of the server blade. Such state transitions can involve the initiation or powering on of the server blade, or the powering down of the server blade following a reset signal, or during other abnormal states during which the operating system is not functioning in the normal manner.

Figure 19 provides an overview of selected components illustrated in Figures 11 and 12 for further illustrating the present invention.

Figure 19 illustrates a server blade 43 containing the blade service controller (BSC) 203. Figure 19 also illustrates a master SSP 74 that forms part of a CSSP 71. Other components are not illustrated for ease of representation.

In an embodiment of the invention, the health of a system module can be monitored during state transitioning, for example when starting or stopping a system module. To this end, the system module, for example a server blade, can output a status signal for predetermined system status points during state transitioning. The monitor module is then operable to start a timer on detecting a first status signal and to reset the timer on detecting a subsequent status signal. The timer can thus be arranged indicate a failed transitioning of the system module in the event that the timer is not reset within a determined period.

In the present example, the system module is a blade server 34 and the monitor module is the SSP 74. In the SSP 74, a timer mechanism 500 is implemented in the processor 301 and memory 305 of the SSP 74. The timer mechanism includes a controller (or control logic) 512, a counter 502, and a timeout alert 504. The control logic 512 can be in the form of computer program instructions held in the memory

305 for controlling the processor 301 of the SSP74. The counter 502 and the timeout alert (T/O A) 504 can also be implemented in computer program instructions held in the memory 305 for controlling the processor 301 of the SSP74. The control logic 512 can be operable to set a predetermined timeout period by adjusting a start point for the counter as represented by the arrows 506. The control logic 512 is also able to monitor a final count reached by the counter 502 on receiving a status signal from the BSC 203 via the link 226. In the event that the counter 502 reaches an overflow value, a timeout signal 510 can be provided to timeout alert logic 504.

Figure 20 is a schematic representation of a time line 600 illustrating various phases during state transitioning for a system module such as a server blade 43.

A status point 602 represents initial power on or reset of the server blade. A power on or reset status signal can be issued from the BSC in response to a received power on signal.

A period 620 represents a period in which initial powering on of the components within the server blade takes place, whereby, for a status point 604, the BSC is able to issue a signal indicating that the system is live.

There further follows a period 622 in which the service processor configures the blade such that, for a status point 606, the boot firmware is able to start booting the operating system for the blade's main processor 192 (see Figure 11). A further status signal can be issued by the BSC for status point 606.

A period 624 represents a period over which the initial boot operations are performed until, at status point 608, the operating system has booted. The BSC 203 can issue a status signal for status point 608 indicating that the operating system has booted.

During a period 630, the operating system is considered to be operating normally, whereby normal operating system heartbeat signals could be output to a watchdog timer as described earlier.

However, a further status point 610 can be reached where a shut down, or panic power off state, occurs. A further status signal can be issued by the blade service controller for this status point 610. The period 632 represents a period in which the blade server 43 is configured such that it can, at status point 612, issue a reset signal. The BSC can be operable to issue a further status signal for status point 612.

Where reference is made in above to a status signal being issued, this is intended to indicate that a status signal can be provided at or in response to reaching the status point concerned. In one embodiment, the status signals are output by the BSC over a management link 622 to the SSP 74. However, it will be appreciated that the status signals could be output or otherwise provided to the SSP by other means, for example by means of special logic circuitry.

The various time periods 620, 622, 624 and 632 could not be monitored by a normal watchdog function as the operating system for the blade server 43 is not operational during those periods. Accordingly, it is during those periods that an embodiment of the present invention enables health monitoring of the operation of the server blade.

In particular, the system module, here the server blade, issues signals at or in response to respective status point being reached, rather than at some predetermined frequency. In different embodiments of the invention, different sets of status signals could be issued.

As illustrated in Figure 20, status signals are issued for status points representative of a power on or reset status point 602; an initial hardware power up status point 604, a start of operating boot status point 606, an operating system booted status point, a shutdown or panic power-off status point 610 and a system reset status point 612.

In various embodiments of the invention, different sets of the periods indicated 620, 622, 624 and 632 could be monitored. Each of these periods could be monitored separately. Alternatively, rather than separately monitoring step 620 and 622, the combined period 626 could be monitored. Likewise, the whole of the period 628 in which the system transitions between the power on or reset state 602 and the operating system booted state at 608 could be monitored as a single unitary period. In other examples different periods could be monitored. For example, other status points that could be used as the basis for monitoring including power on self test start and power on self test end.

In the present embodiments of the invention, the various periods are monitored by means of a resettable timer, formed by a counter 502 and configured in the SSP 74, that is reset on receipt of a subsequent status signal occurring before overflow or underflow of the counter 502 occurs, as will be explained in more detail in the following.

Figure 21 provides a flow diagram illustrating an example of operation of the present invention.

In step 700, and initial timeout value is set by the control logic 512 for the counter 502. This can be achieved by setting, for example, an initial count value from

which the counter counts down towards zero. The initial count value can be determined in various ways, examples of which will be described later.

At step 702, the control logic 512 is operable to monitor the signal line 226 from the BSC 203 waiting for a status signal.

At step 704, when a status signal is received, the control logic is operable to set the timer by resetting the counter to a predetermined count value representative of the initial timeout period.

At step 706, the counter 502 is clocked (e.g., decremented).

At step 708, a test is made as to whether the timer has underflowed as a result of decrementing the counter.

If it is determined at step 708 that the counter has underflowed, then an error is signaled at step 710 by outputting a timeout signal 510 to timeout alert logic 540.

Alternatively, if the timer has not underflowed, the control logic 512 is further operable in step 712 to monitor for receipt of a status signal. If no status signal is received, control reverts to step 706 and the counter 502 is clocked once more.

Alternatively, if a status signal is received, it is then determined in step 714 whether the status signal is the last status signal in a sequence of status signals. For example, if the SSP 74 is operable to monitor each of the transitions between the states 602, 604, 606 and 608, separately, as represented in Figure 20 by the periods 620, 622 and 624, respectively, then four status signals in the series will be expected. Alternatively, if the service processor 74 is merely operable to determine a single period 628 represented by status signals 602 and 608 (where, for example, signals 604

and 606 are not issued by the BSC) then only two status signals are expected. The control logic can be preprogrammed to know the number of status signals to be expected for a given implementation.

If the status signal received is the last status signal, then the monitoring process ends at step 718.

Alternatively, if the status signal received is not the last status signal, then at step 716, the timer is reset by resetting the counter 502 to an initial value and control passes back to step 706.

When resetting the timer at step 716, the control logic can be operable to use different timeout periods between respective sets of status signals, depending on the desired timeout period. Accordingly, a different start value could be used for the initial timeout period compared to the start value for a subsequent timeout period.

As regards the operations to be performed in step 718, this will depend on the status represented by the final status signal. Thus, for example, if a final status signal is representative of status point 608 in Figure 20, this will mean that the operating system is then fully operational, and that a normal watchdog function using the operating system of the server blade may be operational. Alternatively, if the final status signal is representative of a reset state at status point 612, then the next step could be that the process is terminated if the blade server is being finally turned off. Alternatively, it could form a transition to status point 602 if the server blade is to be restarted.

In the above description, it will be noted that reference is made to the counter 502 being set to an initial value and then counting down to an underflow value. Here reaching the underflow value is indicative that the server blade has not reached the

desired status point within the timeout period represented by the counter, and that remedial action is required. Of course, in another embodiment, the counter could be configured as a count up timer where a count is made from an initial count value to an overflow value. Further, rather than having a variable start position and a fixed overflow/underflow value, it will be appreciated that the counter could be configured with a fixed start position and a variable underflow or overflow value. Also, rather than adjusting start and finish values for the counter, the timeout period could also be varied by adjusting the clocking period applied to the counter. Indeed, it will be appreciated that there are many alternative options for timing periods within which an further status signal is expected.

As indicated above, the time out period between respective status points can be different depending on an expected time for the server blade to transition between the status of the two status points. Typically, the time out period will be set as a significant multiple of the actual time expected (e.g. 1.5, 2, 5, etc. times the expected period) in order that minor variations in the actual time to transition between those states does not result in an error condition being indicated.

In a system such as the system described above where the overall system is intended to be highly reliable, but that the individual components, such as the individual blade servers, need not be so highly available, it may be necessary to re-set the server blade from time to time to resolve transient faults.

During the successive resetting operations, it is possible for the SSP 74, as a monitor system, to refine the expected time out period based on actual transition timings experienced during previous resets. Accordingly, the control logic can, for example, be operable to monitor the final count value on a reset occurring (as represented by 508) to record that final count value, and to use the final count value to

determine a new start value for a subsequent instance of the particular transition concerned.

This is represented schematically in Figure 22.

Accordingly, in Figure 22, consider, for example, a situation where the period between status points 602 and 608 is monitored, namely the period 628.

In the first instance, the SSP control logic could be set to apply a very large value to the timeout period, for example, as represented at BLADE BOOT 0, an initial default value of one year could be set.

Optionally, the user could be given the opportunity to modify the timeout period. For example, the user could reset the initial timeout period, as represented at BLADE BOOT 1 to an initial conservative value of five hours.

Consider then that a first reset occurs and, as indicated at BLADE BOOT 1, an actual boot time of 1.5 minutes occurs, as represented by a final count value of the counter 502. The control logic 512 could be responsive to this count to reset the timeout period, for example to 1 hour.

Consider then that a further reset occurs and, as indicated at BLADE BOOT 2, an actual boot time of 1.2 minutes occurs, as represented by a final count value of the counter 502. The control logic 512 could be responsive to this count to reset the timeout period, for example to 10 minutes.

This process could then continue for successive boot operations.

Alternatively, it could continue until some predetermined relationship between the historical actual boot time and the timeout time is established.

The timeout time to be applied could, for example, be determined from tables held in the control logic based on a sequence of the actual boot times. Alternatively a dynamic algorithm could be used, whereby a new timeout time is determined based on a history of boot times. Alternatively, a predetermined multiple (e.g., x1.5, or x2, or x5) of a previous, e.g., the last, boot time could be used to determine a new timeout time to be applied. Of course, where reference is made to "boot times", this could be applied to any timed period.

Also, changes in the configuration of the server blade could be taken into consideration. For example, if a configuration is changed, the timeout period could be reset to an initial default value or an initial conservative value.

In each case, the timeout period should be determined to exceed an expected maximum time to a subsequent status signal assuming a healthy system module.

In the example described above, the SSP 74 is the entity responsible for distributing and/or setting the configuration of the components of the shelf. Accordingly, it will know the pertinent parameters for setting the timeout period, such as whether a network or a disk boot is to be employed for the server blades concerned. Accordingly, the SSP 74, acting as the monitor module, should be able to determine the appropriate timeout period for a system module (e.g., a server blade 43) to be applied dependent upon the configuration of that system module.

Alternatively, or in addition, the system module could be operable to inform the monitor module of a determined period to be applied for the timer. For example,

the BSC 203 in the blade 43 could be provided with configuration information in the EEPROM 316 that it communicates to the SSP 74.

Alternatively, or in addition, the BSC 203 in the blade 43 can be operable to provide the monitor module with details of the configuration of the system module, and the control logic 512 in the SSP 74 can be operable to use the configuration information to compute a determined period to be applied for the timer.

Alternatively, or in addition, the control logic 512 in the SSP 74 can be operable to interrogate the BSC 203 in the blade 43 to determine details of the configuration of the system module and then to use the configuration information to compute a determined period to be applied for the timer.

Figure 23 is a flow diagram illustrating an example of operation of another embodiment the present invention. In this example, a situation where the period between status points 606 and 608 is monitored, namely the period 624 is considered. In this example, the time to be monitored by the timer mechanism 500 is modified dynamically during its operation to take account of changing estimates of the expected boot time.

In step 800, the BSC 203 of the blade 43 issues a signal indicative of the power on or reset.

In step 802, the control logic 512 of the SSP 74 determines an initial time for the timeout period. This could be determined in any of the ways discussed above. For the sake of an illustration, let us say that a timeout period of 10 minutes is chosen. The control logic 512 could then set the counter 502 appropriately.

In step 804, the BSC determines whether the boot is to be performed from a disk (disk), or whether the boot is to be performed over a network (net). In the latter case, it is to be assumed that a longer time would be needed because the bandwidth would be lower for loading the necessary code.

If it is determined in step 804, that the boot is to be performed from a disk, then in step 806, the BSC signals this to the control logic 512 of the SSP 74.

In step 808, the control logic 512 of the SSP 74 then recalculates the timeout period. It could, for example, determine that given that the boot is from disk, and given that some time has already elapsed, that a remaining timeout period of 5 minutes is sufficient. This determination could be made on the basis of data held in timing tables 514 in the control logic. The control logic 512 could then adjust the count of the counter 502 appropriately.

In step 810, the BSC 203 can be operable to probe the devices and configuration of the blade 43. It can then report the results of this to the control logic 512 of the SSP 74.

In step 812, the control logic 512 can be operable to determine estimates of times to boot based on the detected configuration (for example as held in the timing tables 514 in the control logic 512). It might then determine, given that some time has already elapsed, that a remaining timeout period of 4 minutes is sufficient. The control logic 512 could then adjust the count of the counter 502 appropriately.

In step 814, it is assumed that the OS booted status point 608, without the timer having timed out, and then the normal watchdog operation for health monitoring can take over.

Returning to step 804, if it is determined in this step, that the boot is to be performed over a network, then in step 816, the BSC signals this to the control logic 512 of the SSP 74.

In step 818, the control logic 512 of the SSP 74 then recalculates the timeout period. It could, for example, determine, given that the boot is over the network and although some time has already elapsed, that a timeout period of 20 minutes needs to be set to allow for the necessary download. This determination could be made on the basis of data held in timing tables 514 in the control logic. The control logic 512 could then adjust the count of the counter 502 appropriately.

In step 820, the BSC 203 can be operable to analyze the download rate and the amount of information to be downloaded. It can then report the results of this to the control logic 512 of the SSP 74.

In step 822, the control logic 512 can be operable to determine estimates of times to boot based on the detected download information (for example based on the timing tables 514 in the control logic 512). It might then determine, for example if the download speed is slower than might be expected, that the timeout period needs to be increased to 30 minutes. The control logic 512 could then adjust the count of the counter 502 appropriately.

In step 824, for example after the downloading of the necessary information from the network, the BSC 203 can be operable to probe the devices and configuration of the blade 43. It can then report the results of this to the control logic 512 of the SSP 74.

In step 826, the control logic 512 can be operable to determine estimates of times to boot based on the detected configuration (for example as held in the timing

tables 514 in the control logic 512). It might then determine, for example, given that further time has already elapsed, that a remaining timeout period of 6 minutes is sufficient. The control logic 512 could then adjust the count of the counter 502 appropriately.

In step 828, it is assumed that the OS booted status point 608, without the timer having timed out, and then the normal watchdog operation for health monitoring can take over.

During the various steps described with reference to Figure 23, the counter 512 will be counting towards the underflow value, albeit that the current count value may be changed from time to time in accordance with the determinations, or calculations of the control logic 512.

As described with reference to Figure 21, although reference is made to the counter 502 being set to a value and then counting down to an underflow value, it will be appreciate that the timing operation could be effected in other ways. For example, the counter could be configured as a count up timer where a count is made from an initial count value to an overflow value. Further, rather than having a variable start position and a fixed overflow/underflow value, it will be appreciated that the counter could be configured with a fixed start position and a variable underflow or overflow value that is adjusted as required. Also, rather than adjusting start and finish values for the counter, the timeout period could also be varied by adjusting the clocking period applied to the counter. Indeed, it will be appreciated that there are many alternative options for determining the varying timing periods.

Accordingly, there have been described techniques for monitoring the health of a system module during state transitioning where an operating system of the system module is not active. A monitor module can be operationally connected to the system

module, which can output a status signal for predetermined system status points during state transitioning. The monitor module can start a timer on detecting a first status signal and can reset the timer on detecting a subsequent status signal, whereby the timer can indicate a failed transitioning of the system module where the timer is not reset within a determined period.

Embodiment of the invention can thus monitor the health of a system where the heartbeats issued at regular intervals from an operation system are not available because the operation system is not currently active. Although embodiments of the present system have been described in the context of a blade server system, it will be appreciated that the invention is not limited thereto, and indeed is applicable to any system where a system module can be monitored by a monitor system.

Accordingly, although the embodiments above have been described in considerable detail, numerous variations and modifications will become apparent to those skilled in the art once the above disclosure is fully appreciated. It is intended that the following claims be interpreted to embrace all such variations and modifications.